A METHOD FOR CREATING NIH DATA TRAINING TABLES WITH REDCAP AND NIH XTRACT

John E. Kerrigan, Ph.D.
School of Graduate Studies – Biomedical Health Sciences
Robert Wood Johnson Medical School
Rutgers University

Sally Lu
Ernest Mario School of Pharmacy
Rutgers University

ABSTRACT

A major pre-award administrative challenge research universities face is turnaround time for generation of high-quality NIH Data Training Tables for NIH training grants (e.g., T32, K12, TL1, KL2, R25s) which are required for training grant submission proposals to the National Institutes of Health (NIH). Universities with dedicated training grant submission offices generally require data preparation following a structured timeline of several months in advance of the grant submission due date, while other universities with less or no dedicated support for training grant submissions use an ad hoc approach. In these cases, department or program administrators may collect the data manually, in Excel or REDCap, or similar manually maintained methods for those tables requested by the specific NIH grant announcement for the relevant participating graduate predoctoral and/or postdoctoral (including clinical) training programs across the university, depending on the training focus and the "participating faculty" provided by the proposed program director (PD/PI) for the grant. We describe an efficient "federated" method of data collection and construction for NIH Tables (2, 4, 5A/B, 6A/B & -8A part III/8C part III) for new and renewal applications by combining the use of REDCap and NIH xTRACT, leveraging the strengths of each.

Keywords:

REDCap; NIH xTRACT; NIH Data Training Tables; NIH Training Grants

INTRODUCTION

The NIH funds an extensive portfolio of training grants, both individual and institutional. These include formal training at various levels spanning high school to senior career faculty. While individual fellowships (e.g., F30, F31, and F32) fund a specific individual, institutional training grants provide structured funding for multi-year training programs funding multiple fellowships per annum dependent on cohort year. Training program foci can range from broad topics to smaller areas of focus or specialization, depending upon the NIH institute or program mission (e.g., compare General Medicine (NIGMS) which is broader in scope to the National Cancer Institute (NCI) with its focus on all areas in cancer research), a specific program announcement(s), and the resources needed for the overall success of the proposed training program.

A standard requirement for NIH institutional training grant proposals is a set of data training tables. Though not all tables may be required for all proposals, they provide key institutional data such as the departments and programs included in the proposed training; the funding and training records of participating faculty; and the size

and robustness of the departments and programs in providing enough of an applicant pool for the proposed period of performance. There are currently eight sets of tables, each requiring different time scales for data ranges: some require current data; others require retrospective data up to ten years. Upon completion, even a reasonably-sized training grant proposal can include hundreds of pages of data training tables, uploaded as a separate attachment for the grant submission. Each set of data training tables for each submission is custom built, (i.e., the composition of the participating faculty and related departments/programs is unique to each training grant submission dependent upon the many varied combinations of participating faculty relevant to the training focus of the grant). Characteristics inherent in each customized set of data training tables for each submission include complexity and time needed for preparation, review, and finalization.

In general, the higher the number of participating faculty and departments/programs to be covered in the tables, the heavier the data load to be collected and the longer the preparation time for the tables. Cornell University published on the challenges associated with detailed compilation of the data training tables in particular relative to Table 5 (Trainee Publications), which have been described and addressed with a computer program developed to 'dynamically' produce Table 5 (Albert & Joshi, 2019). The University of Michigan Office of Graduate and Postdoctoral Studies over the years has developed a home-grown database that stores the data associated with most of the data training tables except Table 5 pulling institutional data associated with the participating faculty from various internal sources (research grants office; research administrators; trainees; departments/programs to name a few) (University of Michigan, 2021). However, as with our approach, the Michigan database is not maintained in real time, Table 5 is not covered, and some data need to be rechecked and/or refreshed at time of grant preparation.

For pre- and post- award administration of NIH training grants, it is advantageous for pre-doctoral and post-doctoral (including clinical) training departments/programs of research universities to optimize and maintain their tracking of faculty mentor and trainee institutional data using a common approach to ensure accuracy and consistent reporting. A shared, curated source of faculty and trainee data is far more efficient and accurate than storing data in numerous separate 'data pools' across the organization.

In a decentralized approach, some organizations create/collect separate pools of data among relevant graduate and postgraduate programs across the university; others employ a centralized approach to their data management, requiring sustained and dedicated resources to maintain. We have developed and utilize a federated data collection approach that includes individual faculty, trainee, department, program, and other institutional sources, collecting the data 'just in time' for each submission, only seeking what has not already been collected or which needs refreshing. Periodically mining primary institutional data sources (e.g, graduate program applicant/entrant data, institution awards) eliminates the need to have the database house more than can be feasibly managed, improves data quality, and eliminates the need to repeatedly request this redundant information from faculty/departments for each submission. Over time, as this database organically grows, our goal is to continuously improve data quality and accessibility for training grant submissions while reducing overall administrative burden. In this paper, we report on our development of a database developed in REDCap (Harris et al., 2019; Harris et al., 2009), coupled with the use of NIH xTRACT for data training tables preparation and submission. The REDCap database hierarchy begins with the faculty member information; their research support (primarily other than NIH grants data for NIH xTract Institutional Support); and their trainees. In addition, we have developed two separate REDCap databases to track predoctoral Applicants & Entrants (Table 6A) by graduate program and postdoctoral Applicants and Entrants (Table 6B) by department/program.

A secondary advantage to this federated approach is strategic. Contemplation of the composition of the participating faculty (NIH Data Training Table 2) is generally a key strategic step in designing the training program and can benefit from an early-stage draft of the NIH data tables, adding a data-driven and evidence-based approach

to the program's content and design while providing support for strategic planning of the training program. Used in this way from the outset, the optimal composition, size, balance of academic rank, and other relevant strengths of participating faculty can be shaped by the proposed program's leadership and used to strategically showcase unique strengths and capabilities throughout the proposal while pre-emptively addressing any perceived shortfalls for which reviewers might have questions during their review

MATERIALS AND METHODS

What is REDCap?

REDCap was created in 2004 at Vanderbilt University (Harris et al., 2019; Harris et al., 2009). It originally supported a small group of clinical researchers who needed a secure data collection tool that met HIPAA compliance standards. REDCap quickly became a go-to method for supporting both single and multi-site research studies. REDCap's developers firmly believed that nobody could know the research as well as the researcher. Therefore, a user-friendly web-based interface was introduced in REDCap to provide the researchers direct access to their data. Minimal background knowledge or technical experience is needed to use REDCap; researchers and administrators can directly manage their own projects whenever and however they wish, through any browser on any device. Additionally, users can create surveys on desired data collection points. REDCap offers intuitive design tools via a fluid user interface that eases the development of data collection instruments, including variable naming, variable types, variable categories, and linkage of variables. For the data training tables, REDCap is primarily used to store data as a database. We do not use the "survey" function; however, this function is available if needed. The "reports" function is a powerful tool in REDCap and can be used to report on all data or combinations thereof in the database. REDCap has a variety of file export options for reports as well.

What is NIH xTRACT?

Extramural Trainee Reporting and Career Tracking (xTRACT) is a module within NIH eRA Commons used by applicants, grantees, and other designated personnel to create research data training tables for inclusion in progress reports (RPPR) and new institutional training grants (see https://www.era.nih.gov/erahelp/xtract) (eRA Commons, 2017). xTRACT utilizes a related module within eRA Commons (xTrain) to include appointment and related data to pre-populate some training data for training grant renewal applications. This includes data related to trainee names, selected characteristics, institutions, grant numbers, and subsequent NIH and other HHS awards. xTRACT also allows manual entry of data not found in eRA Commons or xTrain. Manually entered information is stored in xTRACT (stored in the research training dataset [RTD]) and can be re-used when preparing subsequent training table submissions.

Compiling the NIH data training tables requires tracking of a substantial amount of faculty member and individual predoctoral student and postdoc data, including but not limited to:

Faculty: Number of pre- and post-doctoral students, number of predoctoral graduated, number of postdocs in training, number of predocs continued, number of postdocs completed, number of postdocs continued in research, NIH Grants, research interest, eRA Commons ID (or Person ID), and rank.

Pre- and post-doc students: Mentor name, eRA Commons ID, pre- or post-doc status, "in-training" status, start date, end date, trainee research topic, faculty eRA Commons ID, publication status and outcomes.

Program data: Applicant and entrant data for graduate programs; program and/or department entrant and applicant data for postdoctoral training proposals, encompassing the recent 5 years of records.

To keep our database up to date, we collect trainee data on a quarterly basis from the following data sources: (1) Directly from the participating faculty only when working on a training grant submission; (2) Faculty research group laboratory websites; and (3) Faculty CVs. We obtain detailed trainee data and outcomes from the following sources: (1) LinkedIn; (2) ORCID; (3) eRA Commons ID; (4) ResearchGate; and (5) Trainee CVs, when available. When working on a training grant submission, we provide a summary to the participating faculty of the data we have for their trainees to minimize the effort on their end for update, requiring a review and an additive refresh of subsequent trainee information in the database.

FACULTY/TRAINEE DATABASE IN REDCAP - UPLOAD TO XTRACT

NIH Table 2 Data and Data Load into xTRACT

NIH Table 2 consists of Faculty Members data, including **Name, Degree, Rank, Primary Department or Program, Research Interest, and Training Role**. In addition to the faculty information, the Mentoring Record (Items 7-12) also must be recorded over the preceding 10-year period for trainees who have been or are currently engaged in research training under the faculty member's primary supervision. This includes current and former predoctorates and postdoctorates and those who have continued in research or research-related careers.

To obtain faculty data not yet in the database (e.g., new faculty hires), we obtain this data from the faculty member or the faculty member's department. Once the faculty data is received and stored within REDCap, a data export that includes the participating faculty data for a particular submission is exported as a CSV file, formatted in Excel, and saved as a tab delimited text file for import into NIH xTRACT. We maintain the faculty data in REDCap with semi-annual updates manually.

NIH Table 4 Data and Data Load into xTRACT

NIH Table 4 contains current research funding of participating faculty members. This data collection includes the faculty member, funding source, grant number, role on project, grant title, and average grant support per participating faculty member. This data is already available in xTRACT (for NIH grants only) via their eRA Commons ID. xTRACT automatically filters exclusions noted in the instructions, e.g., awards on no-cost extension are excluded and only includes those grants where the faculty member has the PD/PI or Project Lead role on the grant. Grants that are not available within xTRACT include foundation and other federal agency awards (e.g., NSF, HRSA). Relevant funding from these sources can be manually entered based on institutional award data within xTRACT that is maintained by the individual building the RTD (see below for additional information). REDCap data upload follows the previous steps as outlined above in the Table 2 data load. xTRACT automatically calculates the resulting average grant funding per faculty member and recalculates as needed until the table (Table 4, illustrated in Figure 1) is finalized.

Figure 1. Fields Used for Faculty Grants and Other Support (xTRACT Table 4)

Faculty Member	Funding Source	Grant Number	Role on Project	Grant Title	Project Period	Current Year Direct Costs
Average Grant Support per Participating Faculty Member						

Trainee Data (Data needed for Table 5 A/B and Table 8A/C)

Table 5A/B describes publications of those predoctoral/postdoctoral in training. For each trainee, it lists the following:

1. Faculty member (mentor)
2. Trainee name
3. Whether the trainee is past or current
4. Training period (date range in mm/yyyy format)
5. Publications

We obtain publication data from PubMed, Faculty CVs or Google Scholar. Tables 5A/B are covered by the "Participating Students" section. It is important to note that only general student data items 1–4 above (not publications) can be uploaded for those with an eRA Commons ID. For students that do not have an eRA Commons ID, one needs to create Person IDs for those students separately within xTRACT.

After the initial trainee data has been loaded into xTRACT, we use the xTRACT interface with PubMed to add the trainee publications manually into each trainee record (Note: If the publication does not appear in PubMed, the publication can be entered manually into the trainee record—see NIH xTRACT use guide for instructions). Separately in the PubMed web interface we have found that simply searching "Trainee First Name Last Name AND Faculty First Name Last Name" in PubMed is sufficient to identify most of the trainee publications with their faculty mentor; however, this approach does not always work with common surnames. We input the PubMed PMIDs into the trainee xTRACT record for their publications. Each publication is curated manually further in selecting the faculty mentor and trainee name per publication entered into the trainee record in xTRACT. While these operations require manual data entry, xTRACT sorts and formats Table 5, saving time overall.

Limitations of Data Loads into xTRACT.

Not all data fields are populated in xTRACT by the tab-delimited text file for the faculty or the trainees; some manual data entry is needed. While the faculty's "Other Grants" data upload populates this data in Institutional Data in xTRACT, these grants need to be manually assigned to the faculty member in "Other Support" in their xTRACT record, and current year "Direct Costs" are recorded separately here as well. Note that grant direct costs change annually per the annual grant budgeting process; however, xTRACT accounts for these changes for NIH grants only.

LIMITATIONS OF THE DECENTRALIZED APPROACH

In the decentralized approach, we created separate pools of data among relevant graduate and/or postgraduate programs across the university de novo for each submission. The administrative burden to copy and paste data into the NIH table (Word or PDF) template, often from an Excel spreadsheet, may result in potential copy/paste errors. In addition, errors are common when uploading the table's PDF to NIH due to unique characters and active hyperlinks that might cause formatting or compliance issues. Relying on a manual process might also cause data to be sorted incorrectly beforehand (e.g., "Commons ID" is needed first before the "Faculty Rank"), resulting in errors in table compilation and additional administrative burden to correct. Additionally, data pulled from a multitude of sources rather than a single curated source raises potential for incorrect and/or incomplete data. Late-stage modification or revision of the tables when created using Excel or Word require painstaking deletions or additions. Re-sorting can increase the likelihood of unintentional errors or formatting snafus. In addition, the worksheets are sometimes discarded post grant submission and the data cannot be reused or further curated for future use.

RESULTS

We utilize a REDCap database coupled with NIH xTRACT for tables preparation, using tab delimited text file transfers into NIH xTRACT to prepopulate collected survey data and working within xTRACT to leverage its interfaces with other components within eRA Commons and PubMed.

We tested this approach with our REDCap database and NIH xTRACT, resulting in finalized NIH Tables 2, 3, 4, 5A/B that took under 12 hours to develop for 15 participating faculty and 98 trainees. However, note this outcome is entirely qualitative and anecdotal. The stated instructions for the NIH data training tables (National Institutes of Health, 2020) approximate that the average administrative burden for data training tables preparation is four hours. However, the broad experience among grants administrators as noted in the Cornell paper (Albert & Joshi, 2019) quoting the NIH Data Tables instructions (National Institutes of Health, 2020) being the preparation time noted by the government is highly underestimated and preparation of these tables can take as long as three months in their experience. On its training grant support office site, the University of Michigan recommends an overall preparation time for a T32 grant to be one year at a minimum with tables preparation at six months in advance of submission (University of Michigan, 2023). In addition, each set of data training tables is very unique, depending on the number of participating faculty, which can range anywhere from 10 to over 100 faculty (e.g. in a CTSA or a dual predoc/postdoc training program) making it challenging to quantitate the tables' preparation time accounting for the different types of data in the training tables and the uniqueness of the training program—which can also be and often is multidisciplinary. We regret that we do not have capacity to conduct a statistically relevant comparison of approaches but encourage the research administration community to do so, using its approach as a benchmark.

Figure 2. Data Workflow Diagram

Faculty/Trainee Data → RedCap Database → NIH xTract
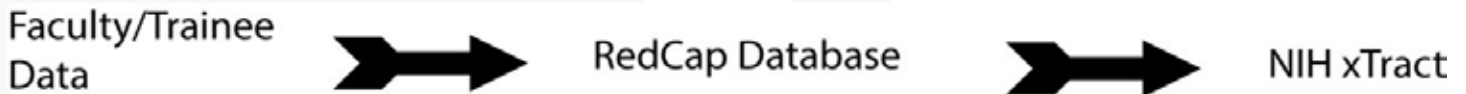
Faculty Data: The faculty data is composed of the following: (a) "NIH Table 2 Participating Faculty" data including their rank; Dept/Program; their role on the grant; research area focus; trainee census; and (b) "NIH Table 4 Research Support"–Faculty NIH research grant support and "Other" support (NSF, Foundation, Other Fed, University). Faculty Data sources include faculty CVs and Office of Research and Sponsored Programs office for the "Other" grants institutional data.

Trainee Data: The trainee data is composed of the following: (a) trainee eRA Commons ID (or Person ID); (b) Trainee type (i.e., Predoc or Postdoc); In training indicator (Yes or No), Training dates (MM/YYYY–MM/YYYY), Publications? (Yes or No) If "No" Why? (New Entrant? Left Program? Change Research Supervisor? Other?); (c) Research Topic; (d) Faculty mentor eRA Commons ID; and (e) co-Mentor eRA Commons ID (optional).

Trainee Outcomes (NIH Table 8 (A or C) part III): The trainee outcomes data is recorded and stored in the REDCap Faculty/Trainee Database manually. The trainee outcomes data for New applications (Tables 8 A&C part III) can be output in CSV format from the REDCap database for entry into the Table template for Table 8 (xTRACT prepares Table 8C part III for new postdoctoral applications; however, xTRACT does not prepare Table 8A part III for new predoctoral applications).

Trainee data & outcomes data sources may include research lab group websites; faculty CVs; LinkedIn (https://www.linkedin.com); eRA Commons ID data if available; ORCID (https://orcid.org); and Doximity (https://www.doximity.com) for medical residents/fellows. For postdoc trainees, data sources additionally may include the human resources records for applicants and payroll system for postdoc entrants in addition to the faculty CV.

Outputs from the REDCap Database: (a) Participating Faculty data upload file for NIH xTRACT; (b) Faculty "Other Grants" data upload file for NIH xTRACT; (c) Trainee data upload file for NIH xTRACT. Trainee Outcomes is output in the form of a CSV file for entry into the Table 8 A/C Word Template.

Note the trainee publications are not stored in the REDCap database. In NIH xTRACT, the publications are obtained manually via PubMed search of the trainee's name and faculty mentor name (using the "AND" logic) as noted earlier. The publications found in the search of the trainee with their faculty mentor from the PubMed search are directly transmitted into NIH xTRACT using the unique publication PubMed PMID (PubMed ID number) of each publication. Note that more complex search of PubMed might be required for "common" trainee names. For example, we have found it useful to add the "Affiliation" or the "Date – Publication" (Date range search available) using the "Advanced" search available in PubMed. Note that publications not found in PubMed can be entered manually into NIH xTract.

Data import into REDCap includes two methods: Method 1 - manual data entry into REDCap by creating a new faculty record; trainee instance or other grant instance; Method 2 - .csv data upload file for multiple records.

APPLICANTS AND ENTRANTS DATABASES (TABLES 6A AND 6B)

Rationale for development of databases for Tables 6A and 6B.

Predoctoral Applicants and Entrants Database for Table 6A

Predoctoral training units can be represented by individual departments OR can be interdepartmental programs (including multi-disciplinary). The Table 6A database in our REDCap database is structured to represent these individual departments and/or programs and store data by academic year going back at least 5 years initially (as required for the data training tables) and maintained per annum. Part I of Table 6A is counts per department of numbers of applicants and entrants (All versus 'Eligible' (i.e., US Citizen or Permanent resident or non-citizen national) per year). Part II is Characteristics (degree school (Prior Institution); GPA; URM %) and the data in Part II is aggregated per year for the case where we have multiple participating departments. The Table 6A database gives the administrator the ability to account for a multitude of "unique" combinations of departments/programs for Part II of Table 6A.

Postdoctoral Applicants and Entrants Database for Table 6B

Postdocs typically reside in academic departments, programs or entities such as institutes or centers within the organization. As with the REDCap Table 6A database, our Table 6B database is structured by department/program and year (academic) going back at least five years. For Part I of Table 6B, the counts requested are counts of applicants/entrants holding the PhD, MD, Dual degree (e.g., MD/PhD; MD/PharmD, etc.) and Other Degree (e.g., DO) per annum in each participating department/program. However, in contrast to Table 6A Part I, in Table 6B Part I this data is aggregated. Table 6B Part II is Characteristics (aggregated) which for postdocs is average number of publications and first-author publications, in addition to "Prior Institution" and URM percentage. Like the Table 6A database, the Table 6B database provides the ability to account for multiple unique combinations of departments/programs.

The REDCap record ID for the Tables 6A and 6B databases represent the participating department/program. The data is stored by department or program in these databases and therefore can be output from REDCap in CSV file format using the REDCap reports tool and selecting the record IDs for the participating department/programs requested. The data retrieved can be aggregated manually as needed for entry either into NIH xTRACT or the NIH table template (Word or Excel).

CONCLUSIONS AND FUTURE DIRECTIONS

Establishing and maintaining an efficient, federated, and shared data compilation method requires strategic and operational planning as well as appropriate resource and knowledge allocation. While NIH xTRACT compiles faculty current NIH grant funding automatically, NSF, foundation, and other grants are best obtained from central university's research administration offices. Development of network data transmission interfaces that feed grants data from institutional data sources (Sponsored programs office) into REDCap for faculty 'other' grants data and human resources data for postdoc data represents another approach to keep these data current and accurate, further reducing administrative burden. Data maintained by Payroll or Human Resources might be useful for postdoctoral applicant and entrant data collection for future grant submissions; navigating a pathway to obtain this data will take time, working with those units which will likely need a better understanding of the context and use of this institutional data. Over time, though, by establishing these pathways and protocols, research institutions should be able to improve accuracy and timeliness for NIH Table 1-8 generation, creating value for those faculty working to increase the number of training opportunities in their institutions. Additionally, an emphasis needs to be placed

on establishing eRA Commons IDs for trainees as early as possible in their training. The method described could be utilized even more effectively if graduate students and postdocs were encouraged to obtain eRA Commons IDs as a matter of common practice, rather than as an afterthought or not at all.

In this paper we have described a more efficient method for managing institutional, faculty, and trainee data, including department/program data for NIH data training data tables preparation using the REDCap database system. Using a singular REDCap database(s) paired with NIH xTRACT has improved data table preparation time, while improving the overall quality of the NIH data tables presentation (xTRACT) and significantly reducing formatting errors, copy/paste errors and administrative burden. We share this in the hope that it will benefit and inspire our colleagues and the research administration community at large.

AUTHORS' NOTE

The work presented does not involve human subjects or animal research. There is no relevant conflict of interest on the part of all authors contributing to this manuscript.

John E. Kerrigan, Ph.D.
Director of Academic Technologies
School of Graduate Studies – Biomedical Health Sciences
Rutgers University
56 College Ave.
New Brunswick, NJ 08901 United States
(848) 932-1525
kerrigje@gsbs.rutgers.edu

Sally Lu
Ernest Mario School of Pharmacy
PharmD Candidate 2024
Rutgers University
New Brunswick, New Jersey, United States
sally.lu@rutgers.edu

Correspondence concerning this article should be addressed to John E. Kerrigan, Director of Academic Technologies, School of Graduate Studies - Biomedical Health Sciences, Rutgers University, 56 College Ave., New Brunswick, NJ 08901, (848) 932-1525, kerrigje@gsbs.rutgers.edu.

REFERENCES

Albert, P. J., & Joshi, A. (2019). Dynamically generating T32 training documents using structured data. Journal of the Medical Library Association, 107(3), 420-424. https://doi.org/10.5195%2Fjmla.2019.401

eRA Commons. (2017). Extramural Trainee Reporting and Career Tracking (xTRACT). National Institutes of Health, NIH.

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, F., Delacqua, G., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software partners. Journal of Biomedical Informatics, 95(103208), 1-10. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)--A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

National Institutes of Health. (2020). Introduction to the data tables. https://grants.nih.gov/grants/funding/datatables/datatables_intro.pdf

University of Michigan. (2021). M-FACTIR: Faculty and Trainee Information Resource.

University of Michigan. (2023). Training program preparation (pre-award). https://ogps.med.umich.edu/training-grant-preparation/