


Research and Applications

A semi-automated pipeline for fulfillment of resource requests from a longitudinal Alzheimer's disease registry

Katelyn A. McKenzie ¹, Suzanne L. Hunt,^{1,2} Genevieve Hulshof,¹
Dinesh Pal Mudaranthakam,^{1,2} Kayla Meyer,² Eric D. Vidoni,^{2,3} Jeffrey M. Burns,^{2,3} and
Jonathan D. Mahnken^{1,2}

¹Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas, USA, ²University of Kansas Alzheimer's Disease Center, Fairway, Kansas, USA and ³Department of Neurology, University of Kansas Medical Center, Kansas City, Kansas, USA

Corresponding Author: Katelyn A. McKenzie, BS, Mail Stop 1026, 3901 Rainbow Blvd. Kansas City, KS 66160, USA; kmckenzie5@kumc.edu

Received 26 March 2019; Revised 21 June 2019; Editorial Decision 11 July 2019; Accepted 22 July 2019

ABSTRACT

Objective: Managing registries with continual data collection poses challenges, such as following reproducible research protocols and guaranteeing data accessibility. The University of Kansas (KU) Alzheimer's Disease Center (ADC) maintains one such registry: Curated Clinical Cohort Phenotypes and Observations (C3PO). We created an automated and reproducible process by which investigators have access to C3PO data.

Materials and Methods: Data was input into Research Electronic Data Capture. Monthly, data part of the Uniform Data Set (UDS), that is data also collected at other ADCs, was uploaded to the National Alzheimer's Coordinating Center (NACC). Quarterly, NACC cleaned, curated, and returned the UDS to the KU Data Management and Statistics (DMS) Core, where it was stored in C3PO with other quarterly curated site-specific data. Investigators seeking to utilize C3PO submitted a research proposal and requested variables via the publicly accessible and searchable data dictionary. The DMS Core used this variable list and an automated SAS program to create a subset of C3PO.

Results: C3PO contained 1913 variables stored in 15 datasets. From 2017 to 2018, 38 data requests were completed for several KU departments and other research institutions. Completing data requests became more efficient; C3PO subsets were produced in under 10 seconds.

Discussion: The data management strategy outlined above facilitated reproducible research practices, which is fundamental to the future of research as it allows replication and verification to occur.

Conclusion: We created a transparent, automated, and efficient process of extracting subsets of data from a registry where data was changing daily.

Key words: reproducible research, dynamic data, research data management, National Alzheimer's Coordinating Center

BACKGROUND

The push for data sharing and increased transparency has been gaining more attention from federal organizations, global organizations, and peer-reviewed journals. In 2003, the National Institute of Health released a statement on sharing research data in which they

endorsed the sharing of research data as vital to furthering the goals of medicine: translating scientific research into “knowledge, products and procedures to improve human health.”¹ In 2012, this concept was echoed by the World Health Organization in a report that stressed the need for international collaboration and public/private sector partnerships to continue the progress of finding solutions for

diseases.² Additionally, in 2016, the *New England Journal of Medicine* reaffirmed their support for transparency in data sharing and encouraged members to allow other researchers to have access to their data.³ While the importance of data sharing and transparency has gained recognition, there are certain situations where the traditional concepts of data sharing and transparency do not apply, thus causing challenges with reproducibility and replicability of research. One such situation is the data collection, storage and distribution of longitudinal, prospective cohort registries.

Longitudinal, prospective cohort studies are becoming increasingly popular for many types of research, including epigenetics, epidemiology, and neurology.^{4–8} Generally, these studies have continual data collection. Additionally, the data collected over time might change with the addition and deletion of certain variables of interest due to evolving research requirements. Thus, these studies are generally considered to have dynamic data as opposed to a clinical trial or experiment with fixed completion times. These changes can occur not only with the addition of new observations, but also with respect to the collection of new variables or similar variables with evolving formats. Such dynamic data issues are more common for registries than clinical trials.

Dynamic data, while invaluable for gathering information and completing numerous research projects, creates many challenges with upholding standard reproducible research principles.^{9–11} These challenges include a lack of a finalized dataset, maintaining the integrity of the data as it is collected and ensuring the accessibility of the data. Moreover, these dynamic datasets can become very large, thus requiring good big data practices.^{12–14} When managing a dynamic database, it is important for these challenges to be overcome in an efficient way so that the dataset can be accessed and utilized for a variety of research projects.¹⁵ While this need has been identified,^{12,16,17} there have been limited methods published in the literature that have outlined an efficient way to overcome these challenges.^{18–20}

The complexities of dynamic data exist in the context of the University of Kansas (KU) Alzheimer's Disease Center (ADC). The Curated Clinical Cohort Phenotypes and Observations (C3PO) database²¹ contains information about clinical, biological, and neural imaging data captured across 935 participants of the KU ADC Clinical Cohort collected since August of 2011. The KU ADC recognized the importance of investigators completing and publishing innovative research, and therefore set a goal to create a seamless process for which investigators had timely access to a current and accurate version of C3PO.²² Prompted by the increase in data requests and variety of data being utilized, this process was required to be flexible enough to curate different types of datasets yet robust enough to decrease data distribution errors.

The primary objective of this article was to describe the process by which the KU ADC enabled investigators to complete reproducible research when using the C3PO dataset. Specifically, this article will identify how the data was curated and stored, how investigators requested data for unique research projects and how subsets of C3PO were generated for the investigators.

MATERIALS AND METHODS

At the time of this writing, the data in C3PO consisted of 1913 variables stored in 15 datasets on 935 patients (Table 1). There were 18 documents describing how the data was collected or generated. The data within C3PO includes various forms of cognitive testing, clinical assessments, blood analysis, imaging studies, and histological

Table 1. Description of all datasets within C3PO that are publicly visible via R2D2

Dataset	Number of variables	Description
UDS	1351	UDS 2.0 and 3.0
Genotype	3	APOE status.
Haplotype	1	Mitochondrial haplogroup.
Blood draw	6	Provides information about the type of blood product to be stored.
Cybrid	2	Indicates the existence of a cybrid line.
Imaging	13	Indicates the type of images available.
Freesurfer imaging	187	Summary measures from MRI imaging.
DXA	33	Body composition.
Cognitive visits	35	Neuropsychological measures unique to KUMC.
CDR visits	23	Data unique to KUMC exclusive of neuropsychological.
Amyloid	1	Quantification of amyloid buildup.
Neuropath	150	Autopsy findings.
Phys function	39	Physical function data.
Physical activity and sleep	55	Physical activity and sleep data.
Milestones	14	Change in participation status.

Abbreviation: UDS: Uniform data set.

reports in addition to basic demographic information. Notably, C3PO contains data on mitochondrial function and distinct aspects of metabolism, such as physical fitness and body circumference measurements.

Data is collected at annual visits, where patients undergo a clinical interview, complete an extensive neuropsychological evaluation and provide biomarkers, such as whole blood samples or imaging. Most of the data is directly entered into Research Electronic Data Capture (REDCap), which is an electronic software designed to support the collection of research data.²³ Data requiring further processing is entered into REDCap as soon as it becomes available. For example, the patient's blood pressure is immediately entered into REDCap while imaging results are entered after analysis by a radiologist. The majority of data collected on the KU ADC Clinical Cohort is also collected at other Alzheimer's Disease Centers; as such, this data is part of the Uniform Data Set (UDS)²⁴ that is maintained by the National Alzheimer's Coordinating Center (NACC).²⁵

Each month, the KU Data Management and Statistics (DMS) Core uploads the UDS to NACC. Quarterly, NACC runs a series of quality control checks, curates the clean UDS data and returns a copy of the curated, clean UDS data to the KU ADC. Site-specific data (ie, data in addition to UDS) is also curated quarterly by the DMS Core and stored with the UDS in C3PO. These quarterly curated datasets are frozen until the next curation, at which time are archived. The data management steps and program for this project was generated using SAS software, version 9.4. Copyright © [2002–2012] SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Investigators seeking to use C3PO data first complete a resource request by submitting a REDCap survey available online. Once approved by the KU ADC Executive Committee, the investigator then meets with a KU ADC member who has substantial clinical

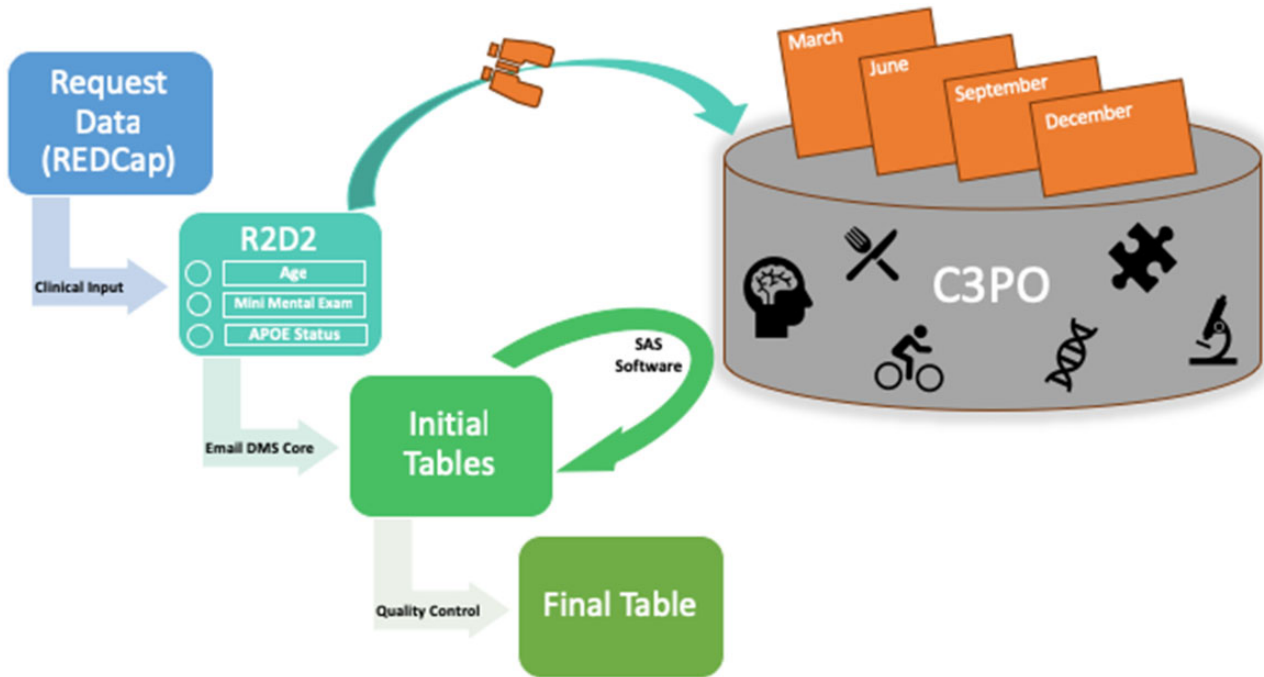


Figure 1. Outline of the fulfillment of KU ADC resource requests from C3PO.

knowledge and experience using the Research Repository Data Dictionary (R2D2),²² a publicly accessible data dictionary for C3PO. The purpose of this meeting is to efficiently identify which of the nearly 2000 variables in R2D2 can be used to answer the research question. R2D2 is searchable so that variable names, descriptions, and formats can be quickly found. The investigators complete their resource request by submitting their variable selections within R2D2. After submission, an email is automatically sent to the DMS Core with the list of requested variables. Within SAS software, this list is utilized to gather necessary information for merging all requested variables, such as the name of the C3PO curated dataset for each variable.

The general strategy for merging all datasets is as follows. First, each C3PO curated dataset, corresponding requested variables, and secondary key information is organized into macro variables so that they can be mutually indexed. Second, an initial table is created that contains all merging information, such as patient ID and visit date, and all variables requested from the UDS. Third, each dataset is added to the initial table, one at a time, by using an iterative macro function and PROC SQL. The macro function creates a separate table for each dataset that is not tied to a particular visit. Please see [Supplementary File 1](#) for the complete program.

After the SAS program generates preliminary tables, the DMS Core examines them for any protected health information to withhold from release and ensures that all requested variables were selected. If the investigator had indicated additional data requirements, such as needing only data on individuals within a specific age range, the DMS Core applies this inclusion criteria. Last, the DMS Core archives the completed requested subset of the C3PO curated dataset by project and date. This step is imperative so that if investigators need to alter their requests, for example to include additional variables requested during manuscript review, the SAS software could access the same curated dataset used in the original extraction. This is extremely important for dynamic data, as pulling data from different quarterly curations may result in modifications,

potentially leading to many downstream negative consequences on prior analyses run. An overview of the methods used in this project can be found in [Figure 1](#) and a brief introductory video can be found in [Supplementary File 2](#).

RESULTS

By using the framework outlined above, 15 data requests were completed in 2018 and 23 data requests were completed in 2017. Over a dozen different departments within the University of Kansas Medical Center (KUMC) used this data for research as well as several other research institutions. The requested datasets supported topics that ranged from how exercise prevented amyloid plaque development to the impact of mitochondrial function on Alzheimer's Disease progression. The average time of data request completion decreased through automation of several steps. Notably, the SAS program took under 10 seconds to compile the requested variables. An example of one such data requests using this framework is provided below. Additionally, a visual demonstration of this process is available in the [Supplementary materials](#).

First, the investigator submitted their project proposal to the KU ADC. Once approved, the investigator worked with the clinical expert to clearly define the data in C3PO that would answer the research question by selecting variables in R2D2. On this same day, the data request was initiated and the DMS Core received an email with the list of requested variables. This list of requested variables was manually input into the SAS program. Before the data subset was sent to the investigator, the DMS Core examined the dataset for any PHI and to ensure the accuracy of the dataset.

DISCUSSION

The data management strategy outlined above resulted in an efficient, effective, and semiautomated process of providing unique

datasets for each project approved by the KU ADC. By having a streamlined procedure for curating C3PO every 3 months, accurate and current, near real-time data was made available to investigators. Importantly, coupled with speed, this framework allowed reproducible research to occur and ensured the protection of sensitive health information.

The novelty of our work is the scale, type and setting. C3PO is a collection of unique datasets that contains 1913 variables collected on over 935 subjects. C3PO contains not only the prospective and longitudinal clinical data from the UDS, but also data on metabolic function, APOE genetics, autopsy findings, amyloid plaques, and magnetic resonance imaging (MRI) scans. R2D2 is both interactive and available to the public, thus facilitating transparency of the data upholding best reproducible research practices. The combination of a dataset and data dictionary such as C3PO and R2D2 that is available to all researchers is less common. As there are limited peer reviewed publications on dynamic data management strategies, the process outlined in this article serves as a solid foundation of how to approach dynamic datasets so that the integrity of clinical research, such as reproducibility, is upheld.

Before the creation of C3PO and R2D2, completing data requests was considerably more time consuming. Often, investigators would use a PDF of the entire data dictionary, which was nearly 150 pages long, to select variables for their project. They would then email the DMS Core this list of variables and the DMS Core would have to verify the exact variables the investigator wanted. Frequently, multiple emails and phone calls between the DMS Core and the investigator would ensue, thus prolonging the time before investigators received the requested data. For example, there are three different variables in C3PO with distinct diabetic classification criteria. Before R2D2, investigators may not have realized there were different diabetic classifications within C3PO. Therefore, if they did not specify which diabetic classification they needed, the DMS Core would not know which variable to select, thus prompting further clarification and delaying the data request process. By using R2D2 and a clinical expert, investigators were better able to identify the appropriate variables for their project because the data standards and formats were publicly available for all variables. Overall, the data request process was expedited with the addition of C3PO and R2D2.

Beyond allowing investigators to more easily select variables, R2D2's structure contains essential information necessary for creating the requested dataset. This information includes the dataset within C3PO that contains the requested variable and the secondary key needed for joining variables from distinct datasets within C3PO. As requests may only require data collected at baseline, data from multiple annual visits, data from a single time point, or a combination of these, the secondary key for each dataset within C3PO is not the same. Therefore, while the sequence of steps to create a requested dataset is dependent on the combination of variables, we were able to automate the merging of variables from C3PO by utilizing SAS software.

The algorithm used in the SAS software utilizes a user-defined macrofunction that systematically and iteratively searches through each dataset within C3PO to select the requested variables. This user-defined macro function allows for situational joining of subsets of datasets that is completed within seconds. While automation of this method required an investment of time to create, it resulted in two major advantages. First, the DMS core could use the same, efficient SAS program to create the datasets ([Supplementary File 1](#)). Previously, the SAS program would have to be manually changed for

each data request. This process was complicated and inefficient because the statistician had to either memorize or look up which datasets contained each of the requested variables and the corresponding secondary key. By using the same, automated SAS program, the DMS Core was able to increase their efficiency and minimize errors. For example, it used to take between 3 and 4 hours to complete a data request but now it takes between 1 and 2 hours. Second, the automated SAS program could identify data management discrepancies, therefore helping with data quality checks. For example, if the data dictionary recorded a variable as being in the incorrect C3PO curated dataset, a note was written to the SAS Log, alerting the DMS Core of this mistake. This is extremely important for ensuring the quality of the data dictionary.

One of the most important aspects of this method is that it ensures the data request process follows reproducible research practices. This was done by both curating the data quarterly and using an automated program. Curating the data quarterly established an effective final dataset and automating the SAS program allowed for replicability. Moreover, this project advanced reproducible research practices by increasing the access and transparency of the data by making R2D2 publicly available and searchable. This helped investigators take more ownership of the data. For example, investigators were able to select the exact variables they needed to answer their research hypothesis and knew the format of this data. Establishing standard reproducible research practices in studies with dynamic data is vital to the future of research in healthcare because if the results cannot be verified, then the credibility of the results decreases.

This process has two weaknesses. First, examining the accuracy of the created dataset and ensuring the privacy of protected health information cannot be automated. While it would be ideal for this step to be automated, it would be nearly impossible to do so because of the importance of protecting patient's rights. Second, timely maintenance of technology and the data dictionary is necessary for maintaining the credibility and security of the data. External factors, such as SAS Software updates, require a higher level of technical effort to maintain this program.

The main strengths of this method include replicability, transparency, efficiency, and accuracy in the context of dynamic data. All of these qualities uphold reproducible research standards. By employing this automated process, the KU ADC is better able to support investigators and resourcefully utilize data the KU ADC cohort has provided. This process is a major step forward not only for reproducibility practices, but also for fostering positive collaboration across many disciplines at a large research institution.

CONCLUSION

We have described a process that allows for reproducible research in a longitudinal clinical cohort with dynamic data. This strategy utilizes quarterly data freezes to ensure that current snapshots of our dynamic data are available to investigators. Additionally, much of this process is automated, thus allowing for the data to be disseminated quickly and efficiently to investigators while ensuring the quality of the data.

FUNDING

This research was supported by NIH grant P30 AG035982 through the National Institute on Aging.

AUTHOR CONTRIBUTIONS

SLH, DPM, KM, EDV, JMB, and JDM contributed to the conception and overall design of C3PO and R2D2 in addition to overseeing data collection. KAM, SLH, DPM, and JDM contributed to the creation and execution of the methods used to facilitate and fulfill data requests. KAM, SLH, GH, and JDM drafted the initial manuscript. All authors were involved in critically revising the manuscript and approved the submitted manuscript.

DATA SHARING

Data contained within C3PO is available upon request pending appropriate scientific and safety review (<https://redcap.kumc.edu/surveys/?s=wQMXHa>). A complete list of variables contained within C3PO is publicly available (<http://r2d2.kumc.edu/ADC/R2D2.jsp>).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST

None to report.

REFERENCES

- National Institute of Health. *Final NIH Statement on Sharing Research Data*. NIH Guide; 2003.
- World Health Organization and Alzheimer's Disease International. *Dementia: A Public Health Priority*. 2012; 112. https://www.who.int/mental_health/publications/dementia_report_2012/en/
- Drazen JM, Morrissey S, Malina D, et al. The importance—and the complexities—of data sharing. *N Engl J Med* 2016; 375 (12): 1182–3.
- Ng JW, Barrett LM, Wong A, et al. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol* 2012; 13 (6): 246.
- Wang SV, Verpillat P, Rassen JA, et al. Transparency and reproducibility of observational cohort studies using large healthcare databases. *Clin Pharmacol Ther* 2016; 99 (3): 325–32.
- Li K, Chan W, Doody RS, et al. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J Alzheimers Dis* 2017; 58 (2): 361–71.
- Fritz NE, Newsome SD, Eloyan A, et al. Longitudinal relationships among posturography and gait measures in multiple sclerosis. *Neurology* 2015; 84 (20): 2048–56.
- Mahmood SS, Levy D, Vasan RS, et al. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* 2014; 383 (9921): 999–1008.
- Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol* 2015; 13 (6): e1002165.
- Laine C, Goodman SN, Griswold ME, et al. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* 2007; 146 (6): 450–3.
- Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. *Am J Epidemiol* 2006; 163 (9): 783–9.
- Wang X, Williams C, Liu ZH, Croghan J. Big data management challenges in health research—a literature review. *Brief Bioinform*. 2019;20(1):156–167. doi:10.1093/bib/bbx086.
- Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics* 2016; 22 (2): 303–41.
- Belle A, Thiagarajan R, Sorousmehr SMR, et al. Big data analytics in healthcare. *Biomed Res Int* 2015; 2015: 370194.
- Mayer-Schonberger V, Engelsson E. Big data and medicine: a big deal? *J Intern Med* 2018; 283 (5): 418–29.
- Anderson NR, Lee ES, Brockenbrough JS, et al. Issues in biomedical research data management and analysis: needs and barriers. *J Am Med Inform Assoc* 2007; 14 (4): 478–88.
- Johnson SB, Farach FJ, Pelphrey K, et al. Data management in clinical research: synthesizing stakeholder perspectives. *J Biomed Inform* 2016; 60: 286–93.
- Nind T, Galloway J, McAllister G, et al. The research data management platform (RDMP): A novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data. *Gigascience*. 2018;7(7):giy060. doi:10.1093/gigascience/giy060.
- Brembilla A, Martin B, Parmentier A-L, et al. How to set up a database?—a five-step process. *J Thorac Dis* 2018; 10(Suppl 29): S3533–S3538.
- da Silva KR, Costa R, Crevelari ES, et al. Global clinical registries: pacemaker registry design and implementation for global and local integration—methodology and case study. *PLoS One* 2013; 8 (7): e71090.
- University of Kansas Alzheimer's Disease Center. *Curated Clinical Cohort Phenotypes and Observations (C3PO)*. 2018. <http://www.kualzheimer.org/researcher-resources/available-scientific-resources.html>. Accessed December 2018.
- University of Kansas Alzheimer's Disease Center. *Research Resource Data Dictionary (R2D2)*. 2018. <http://r2d2.kumc.edu/ADC/R2D2.jsp>. Accessed December 2018.
- Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; 42 (2): 377–81.
- National Alzheimer's Coordinating Center. Uniform Data Set (UDS). 2015. https://www.alz.washington.edu/WEB/forms_uds.html.
- National Institute on Aging. National Alzheimer's Coordinating Center (NACC). 1999. https://www.alz.washington.edu/WEB/researcher_home.html.