SAS® GLOBAL FORUM 2020 #SASGF

Paper 4152-2020

Using SAS® and R Integration to Manage and Create a Multilevel Complex Database: A Case Study of the U. S. Department of Defense Military Health System Data Repository

Akhtar Hossain, University of South Carolina, Columbia; Nikki R. Wooten, University of South Carolina, Columbia; Laura A. Hopkins, Kennell & Associates, Incorporated

ABSTRACT

As the most established statistical analysis system, SAS® is widely used by researchers and programmers in diverse fields. R, an open-source statistical programming language, has widespread usage among small and medium enterprises as well as both researchers and programmers. Despite widespread use, both SAS® and R have advantages and limitations. Consequently, integrating SAS® and R provides an efficient solution to practical data management and analysis problems. Using the U. S. Department of Defense's Military Health System Data Repository, we present a case study integrating SAS and R to manage and analyze complex, multi-sourced administrative and medical claims databases. SAS code to utilize the R interface within SAS is also provided.

INTRODUCTION

Historically, SAS® has been the statistical analysis system most researchers and enterprises rely on for data management, analyses, and reporting. However, customizations of SAS programs and macros for case-specific data manipulation, extraction, analyses, and graphics are difficult and pose challenges to many researchers. This difficulty hinders efficient data management, manipulation, and analysis tasks for researchers without advanced SAS programming knowledge. On the other hand, R, an open-source matrix-based object-oriented statistical programming language, has advanced significantly in the past two decades to become one of the most widely used languages in the data sciences and is a powerful and compelling alternative to SAS for many users.[1]

In many instances, researchers enjoy the respective advantages of SAS® and R, but are limited by their drawbacks as well. To take advantage of the benefits of R, SAS has incorporated the R interface within the SAS system. Starting with SAS® 9.3, Windows and Linux users with the SAS/IML® license can use the R interface within SAS to call R functions and can transfer data between SAS and R.[2,3] Years after introducing this SAS - R integration, the usability and efficiency that can be gained by researchers remains relatively unknown and few papers have been presented and published on the basic usage of this integration. In this paper, we present a case study demonstrating the utility and efficiency of the integration of SAS and R for management and analysis of complex, multi-sourced databases using the U. S. Department of Defense’s Military Health System Data Repository (DoD MDR).

U. S. DEPARTMENT OF DEFENSE’S MILITARY HEALTH SYSTEM DATA REPOSITORY

This case study stemmed from the data management and analysis challenges experienced in the Army Warrior Care Project (AWCP),[4] which uses the U. S. DoD MDR to examine behavioral health problems and service utilization among Army service members who were assigned to Warrior Transition Units after returning from Operations Enduring Freedom, Iraqi Freedom, and New Dawn deployments between FY2008−2015. The MDR is the most comprehensive and reliable source of Military Health System data available to researchers.[5] It is a database integrated from multiple sources and different types of information, including records on all health care events, historical beneficiary data, TRICARE coverage information, military service records, demographics, and clinical data. The comprehensive nature of the DoD MDR involves the sourcing data from multiple origins, at multiple levels (person-level, prescription-level, and visit-level), and at multiple timepoints (fiscal years 2008-2015, pre-deployment, deployment, and post-deployment),[4] which made efficient management, extraction, and analysis of data to accomplish AWCP study aims very challenging.

The AWCP utilizes data elements from several domains of the MDR, including demographic and military service record data; deployment characteristics; behavioral health diagnoses and service utilization; prescription drug use; in-theatre behavioral health diagnoses and service utilization; physical injuries; and Warrior Transition Unit characteristics.[4] These data elements are from 10 different sources, contains data at three different levels (person-level, prescription-level, and visit-level) and most importantly includes approximately 140 million data lines on more than 1,000 variables and 900,767 soldiers. The AWCP project received data in three larger domain files for person-level data (~1 million data lines), prescription drug data (~40 million data lines), and visit-level data containing inpatient, outpatient, and emergency department visits and medical diagnoses (~100 million data lines).

The AWCP used SAS as the primary data management and analysis software for person-level, prescription-level, and visit-level data. However, the size, complexity, and need for creating customized variables for different phases of the deployment cycle (e.g., pre-deployment, deployment, post-deployment), military service periods (e.g., pre-September 11th), and time periods (e.g., fiscal year 2008)[4] made data manipulation and management very challenging using only SAS. Using the object-oriented programming and rich collection of R packages in combination with SAS provided an efficient solution. Thus, data management and extraction processes were initiated with SAS and the R interface integration in SAS was used to initiate and accomplish complex data processing, extraction, and analysis in SAS. We present sample codes and coding algorithms in the next section and Figure 1.

SAS AND R INTEGRATION TO MANAGE AND ANALYZE MDR DATA

To use R functions within SAS, the SAS workspace server and R must be downloaded (http://cran.r-project.org/) and installed on the same computer. According to the SAS support note (https://support.sas.com/rnd/app/studio/Rinterface2.html), SAS 9.2 or later needs to be installed and IML Studio version 3.3 or higher is also required. However, PROC IML in SAS 9.22 does not support R versions starting with 2.12. Thus, it is advised to use R 2.11.1 instead. SAS 9.3 does not have this limitation (http://support.sas.com/kb/42/079.html).

By default, R functions cannot be used within SAS. This is specified by the systems option NORLANG. To access R, SAS needs to be started with the RLANG Option (C:\...\sas.exe -RLANG). One can confirm that R is installed and is accessible within a SAS session using a PROC OPTION step

```
PROC OPTIONS OPTION = RLANG;
RUN;
```

If SAS and R are running properly, the above statements should return the message “RLANG: Support access to R language interfaces” in the log section. Once SAS and R are installed and setup, both programs work well together and it becomes very easy to transfer data back and forth between them. Figure 1 presents a flow diagram of the algorithm used in the AWCP case study. After reading the original datasets into SAS, some initial data cleaning steps were performed within SAS. The datasets were then exported to R and R functions were called within IML to create new variables and dataframes. See SAS code below.

```
PROC IML;
RUN ExportDataSetToR(“Sdata_per”, “Rdata_per”);
RUN ExportDataSetToR(“Sdata_vis”, “Rdata_vis”);
:
SUBMIT / R;
   :
   The R Functions and Code for Necessary Calculations
   :
ENDSUBMIT;
QUIT;
```

Once R completed computational tasks and created the required dataframes, the dataframes were imported back into SAS for further statistical analyses per AWCP study aims and data analysis plans. Programming codes are provided below for further understanding.

```
PROC IML;
RUN ImportDataSetFromR("Sdata_per_vis", "Rdata_per_vis");
:
QUIT;
```

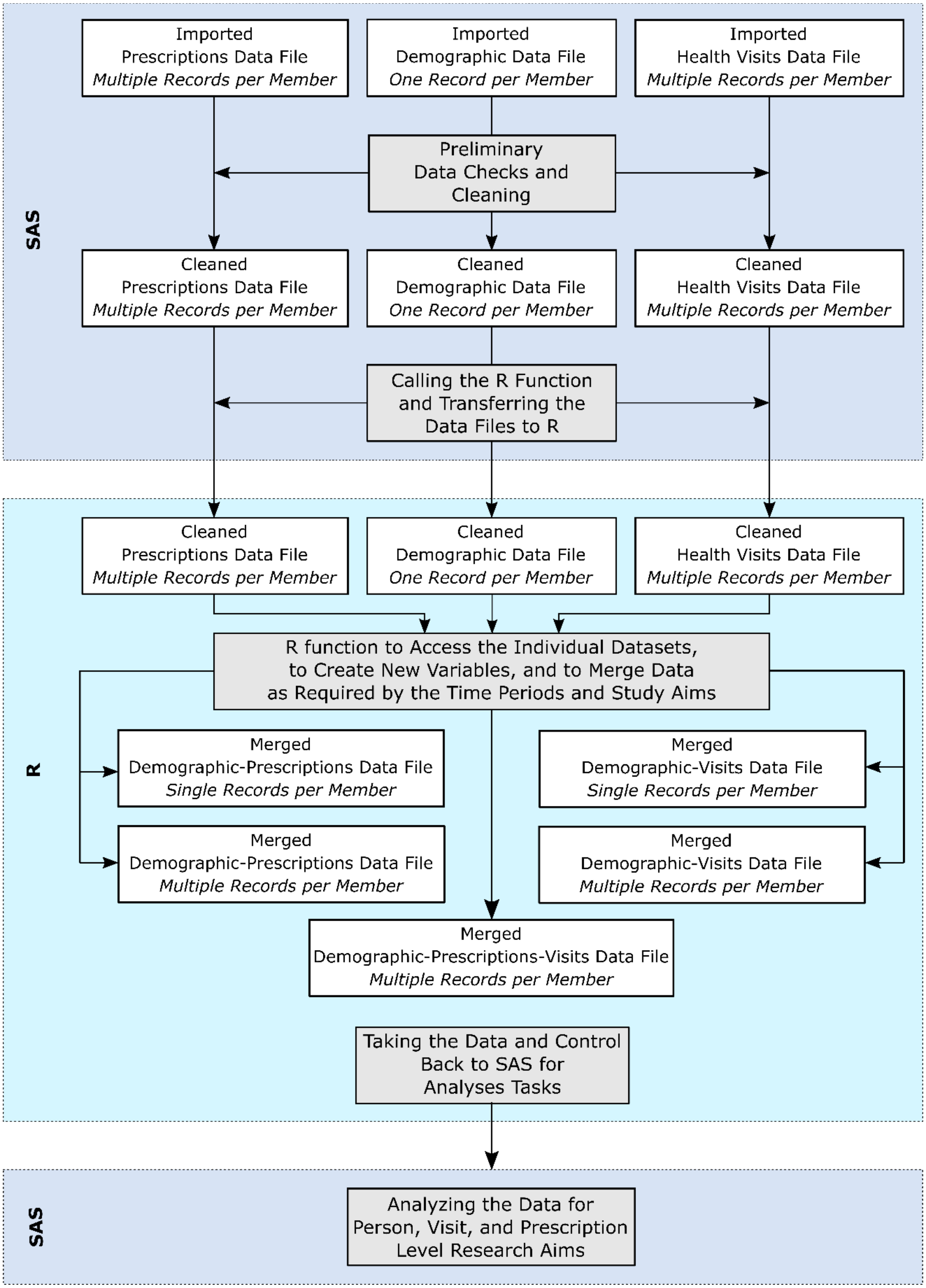
SAS
Imported
Prescriptions Data File
Multiple Records per Member
Imported
Demographic Data File
One Record per Member
Imported
Health Visits Data File
Multiple Records per Member
Preliminary
Data Checks and
Cleaning
Cleaned
Prescriptions Data File
Multiple Records per Member
Cleaned
Demographic Data File
One Record per Member
Cleaned
Health Visits Data File
Multiple Records per Member
Calling the R Function
and Transferring the
Data Files to R
R
Cleaned
Prescriptions Data File
Multiple Records per Member
Cleaned
Demographic Data File
One Record per Member
Cleaned
Health Visits Data File
Multiple Records per Member
R function to Access the Individual Datasets,
to Create New Variables, and to Merge Data
as Required by the Time Periods and Study Aims
Merged
Demographic-Prescriptions Data File
Single Records per Member
Merged
Demographic-Visits Data File
Single Records per Member
Merged
Demographic-Prescriptions Data File
Multiple Records per Member
Merged
Demographic-Visits Data File
Multiple Records per Member
Merged
Demographic-Prescriptions-Visits Data File
Multiple Records per Member
Taking the Data and Control
Back to SAS for
Analyses Tasks
SAS
Analyzing the Data for
Person, Visit, and Prescription
Level Research Aims

Figure 1. SAS, R integration algorithm for managing and analyzing complex, multi-level, multi-sourced datasets from the U. S. Department of Defense's Military Health System Data Repository

CONCLUSION

In research projects involving complex, multi-level, multi-sourced data, data management and analysis algorithms can be efficiently processed by taking advantage of the SAS - R integration. This paper presented a case study of using the integration of proprietary (SAS) and open-source (R) statistical software systems for the efficient management and analysis of a large scale, longitudinal, multi-sourced database created using data elements from the U. S. Department of Defense's Military Health System Data Repository. This case study demonstrates the utility and efficiency of the SAS - R integration for complex data manipulation and encourages researchers and programmers to take advantage of the powerful and efficient statistical software integration in various research settings.

REFERENCES

1. Using the R interface in SAS® to call R functions and transfer data https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3556-2019.pdf
2. Working with both proprietary and open-source software https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3089-2019.pdf
3. SAS and open source: Two integrated worlds https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3415-2019.pdf
4. Wooten NR, Brittingham JA, Hossain A, et al. Army Warrior Care Project (AWCP): Rationale and methods for a longitudinal study of behavioral health care in Army Warrior Transition Units using Military Health System data, FY2008−2015. *International Journal of Methods in Psychiatric Research*. 2019; e1788. https://doi.org/10.1002/mpr.1788
5. Guide for DoD researchers on using MHS data, Office of the Assistant Secretary of Defense for Health Affairs (OSAD(HA)), Defense Health Agency, Human Research Protection, August, 2012. https://health.mil/Reference-Center/Publications/2012/10/10/Guide-for-DoD-Researchers-on-Using-MHS-Data

FUNDING

This study was funded by a National Institute on Drug Abuse Mentored Research Scientist Career Development Award (#K01DA037412; PI: Dr. Wooten).

ACKNOWLEDGMENTS

Aggregate data files for this study were created by Laura Hopkins, MS, Kennell & Associates, Incorporated. Diana D. Jeffery, PhD, Clinical Services Division, Defense Health Agency is the program official who sponsored access to data from the U. S. Department of Defense Military Health Systems Data Repository. The opinions and assertions herein are those of the authors and do not necessarily reflect the official views of the U. S. Department of Defense, National Institute on Drug Abuse, or the National Institutes of Health.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Akhtar Hossain
Department of Epidemiology and Biostatistics
University of South Carolina, Columbia, SC
mhossain@email.sc.edu

Nikki R. Wooten, PhD, LISW-CP
College of Social Work
University of South Carolina, Columbia, SC
nwooten@sc.edu

Laura A. Hopkins, MS
Kennell & Associates, Incorporated
Falls Church, VA
lhopkins@kennellinc.com