Taking The Data Validation To Next Level: Automating Data Validation Using CDASH-Standardized Global eCRFs

Shreya Rachuri, MSc; Mary O'Dwyer, MRP, CCRP; Kathe Douglas, BA; Leilani Logan, MSc; Josh Tewell, BA; Saianand Balu, MS;
Carrie Lee, MD, MPH; J. Kaitlin Morrison, PhD; Erin Crecelius, MA
[Lineberger Comprehensive Cancer Center]

Background

A high volume of clinical trial data is generated daily in Investigator-Initiated trials(IITs) at Lineberger Comprehensive Cancer Center. Data management staff then ensure the data's accuracy, reliability, and consistency and prepare high-quality datasets for safety and efficacy analysis by biostatisticians. Hence it is essential to have the data validated from the initial stages of the clinical trial to avoid risks to patient safety and data quality. Data validation is a tedious and time-consuming process highly susceptible to human error, resulting in lower data quality. Therefore, there is a necessity for automating the data validation process, which is time-saving and more efficient than manual data validation. This automation helps in understanding multivariate data relations.

Goals

This project aims to implement an automated data validation process using CDASH-standardized global Electronic Case Report Forms (eCRFs) to maintain data quality and integrity by automatically detecting non-compliant data. Further goals include:

- Expediting data review by directly reviewing the targeted data
- Increasing accuracy of the data validation by eliminating human-prone errors and increasing query volume
- Decreasing the programming time by standardizing the checks for global forms used across all the studies and reusing the same code
- Developing a user interface to execute the data validation programming checks by entering the study reporting parameters that will run the SAS program and output the data into the study folder

Solutions & Methods

For each study, the Clinical Data Management Associate (CDMA) provides the Data Validation Plan (DVP) to the programmer in an Excel sheet. The automation uses the Statistical Analysis System (SAS) programing language, whereby the checks within and across the eCRFs are programmed into separate SAS code files. After executing the programs, data issue reports are generated as Excel workbooks containing one sheet corresponding to each eCRF. Each Excel sheet contains tables of observations where issues are color code and titles describe issues programmed. The CDASH eCRF Global library enables a standardized SAS program file that can be used across all studies.

Fig1: Flow Diagram Of Events In Automating Data Validation

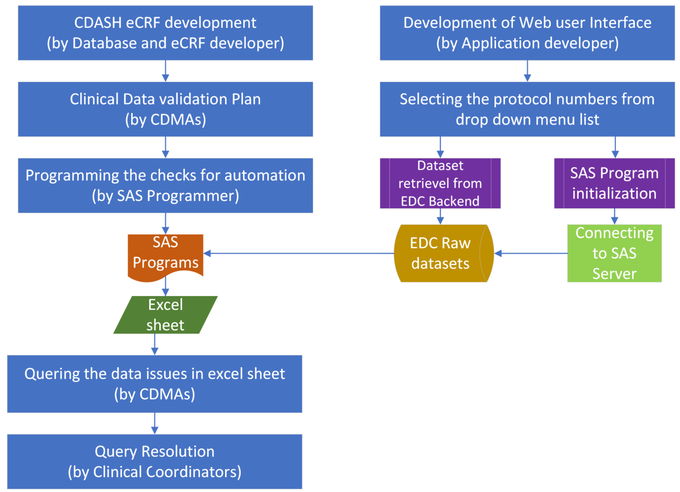
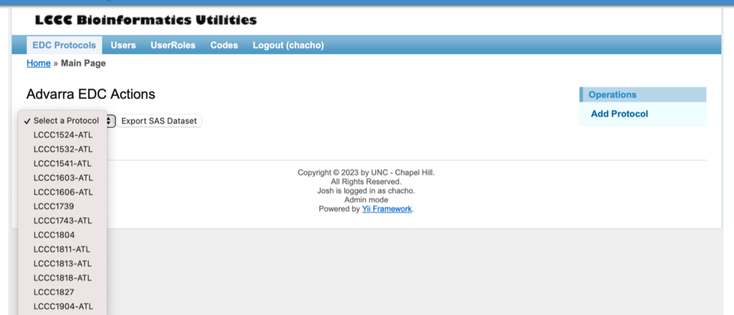


Fig 2: DVP Spreadsheet Prepared for SAS Programmer. All Variables for All Forms are Listed. The CDMA Indicates How Critical The Form Data is And Validation to be Programmed

Form Display Nam	Variable Name	Critical	Manual Check
Medical History	MHCAT	Medium	IF MHTERM_DIAG not null, MHCAT required
Medical History	MHCATOTH	Medium	If MHCAT = Other, MHCATOTH required.
Medical History	MHEDTTYP	Medium	If MHEDTTYP = Diagnosis, then MHTERM_DIAG, MHCAT, MHLOC requird
Medical History	MHENDAT	Medium	If MHTERM not null, required field
Form Display Nam	**Variable Name**	**Critical**	**Manual Check**
Adverse Events	AEACN	High	Should not be null
Adverse Events	AEACNOTH	High	Should not be null
Adverse Events	AECAT	High	Should not be null. If AECAT = Prior to screening, AEREL/AEREL2/AEREL3/AEREL4 should = Not Applicable
Adverse Events	AEENDAT	High	If AEENDAT has date, then AEOUT should be Recovered/resolved Recovered/resolved with sequelae

Fig 3: User Interface to Select Protocol From Drop Down Menu to Connect to The EDC Backend. The Connection Initializes The SAS Program to Generate The Validation Reports



Outcomes

The data review process is faster and easier as all the targeted data is in one place. We will collect metrics to assess the number of discrepancies and data review time using the new process. Automated results made a significant difference when there were many records where the issues could be easily missed in manual reviews. This new process also increases frequency of clinical data reviews. The creation of cross-form checks enabled the assessment of multivariate data relationships. Evaluating standard data checks and queries to streamline the eCRF build increased the database build efficiency. Reusing the code for standardized checks for global forms has decreased the programming time by approximately 50-60%. Success is contingent on improving the data review time.

Fig 4: Validation Outcome in Excel Sheet Listing The Data Issues And Missing Data is Color Coded for Immediate Identification.

If MHTERM_DIAG not null, MHCAT required
If MHCAT = Other MHACTOTH required
if MHTERM not null, mhstdat and mhendat required

Medical History Event Type	Medical History Term	Cancer Diagnosis	Other specify diagnosis	subtype	other specify, subtype	start date	end date	Number of relapses
Current Diagnosis	a	Colon	a	Abdomen	a	15Oct2021	14Oct2021	0.0000000000
General Medical History Event	Anaemia					04Oct2021	.	.
Episode	abc					26Oct2021	26Oct2021	.
						.	.	.

Lessons Learned & Future Outcomes

Some checks cannot be automated and require CDMA review. We successfully standardized the checks for forms, wherein we reduced programming time by coding individual reusable forms common for global studies. New coding will only be required for less common study-specific forms. Phase two of the project will aim to build and host a website where authorized users can choose an Advarra EDC protocol from a list. Upon execution, it connects to the EDC backend and downloads the SAS datasets and then builds a SAS initialization program file. This SAS file has the program that generates the validation reports. Upon invocation of SAS on web application server and executing the SAS initializing file, validation reports are generated. Once all this process is built, Validation reports can be run at any time without the involvement of SAS Programmer.