PharmaSUG 2019 - Paper PO-108

Developing Analysis & Reporting Standards For Pharmaco-Epidemiology Observational Studies

Bo Zheng, Merck & Co., Inc., Upper Gwynedd, PA, USA

Xingshu Zhu, Merck & Co., Inc., Upper Gwynedd, PA, USA

ABSTRACT

In the blossoming world of modern data analytics, there is an unfulfilled need for standardization within real world evidence (RWE). Lack of standardization often leads to longer and more frustrating program development cycles. This paper discusses our experience with developing standards across RWE primary data collection (PDC) studies. We developed a process for Pharmacoepidemiology PDC studies by standardizing variables based on existing CDISC conventions, developing data quality review tools, and creating a set of modular macros that creates a unified deliverables package. Having a core standardization process in place is beneficial for reducing the time it takes to identify and resolve data quality issues and getting deliverables to customers. As RWE continues to expand, implementing standards provides a unique opportunity to help guide its path towards even greater acceptance within the scientific community.

INTRODUCTION

The Food and Drug Administration (FDA) defines RWE as “the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of real-world data”[1]. In the rapidly blossoming world of modern data analytics, there is an unfulfilled need for standardization within the RWE area. Lack of standardization often leads to longer and more frustrating program development cycles; this problem is becoming even more prevalent as acceptance and demand for RWE analysis for clinical development programs continues to grow.

We took a practical approach to standardization to meet our unique needs in Pharmacoepidemiology (PE) that included the following.

- Creating a formal Data Quality Review (DQR) process that involved collaboration between statistical programmers, epidemiologists, and data vendors to clearly define roles and responsibilities of each individual contributor and where those roles overlap.
- Outlining the Data Review Categories (DRC) by the epidemiologists with input from statistical programmers to cover a wide range of data quality related issues that have come up in real world data.
- Defining the standard variable names, creating a SAS® toolset to map raw variables to standard variables names to simplify the initial data import process and to facilitate the development of standard macros for tables, figures and listings (TFL).
- Writing a comprehensive execution resource that covers the entire DQR process, available programming tools, and how to use them to detect, document, and remediate data quality issues.
- Streamlining a large collection of legacy TFLs into a unified deliverables package that can be created with standard macros in SAS® through a joint venture by statistical programmers and epidemiologists.

THE DATA QUALITY REVIEW PROCESS

Defining the complete DQR process was an important first step for assessing and focusing on the individual and collaborative roles of the data suppliers (e.g. vendors and academic collaborators), the epidemiologist, and the statistical programmer.

The DQR process was defined with the following components in mind.

- Statistical programmers' primary responsibility in the process was to develop and test DQR programs and to create DQR process-compliant datasets and reports.
- Statistical programmers also work with the epidemiologist to help negotiate terms in the vendor contract, such as the data dictionary and the periodic receipt of data.
- To kick off the start of each new PE project, the statistical programmers and epidemiologist will collaborate to:
    1. Identify the key variables and data points for DQR.
    2. Define which TFLs are needed for the final report.
    3. Develop the programming specifications for the analysis dataset and deliverables.
- After the raw data is delivered through a compliant process and the initial DQR rounds are completed, any data issues discovered are recorded in the DQR Issues Log and sent to the epidemiologist for review.
- Data entry or collection errors that cannot be resolved internally will be sent back to the vendor, and another delivery of raw data will be requested. The updated dataset also will go through the same DQR process, checking to see if existing issues were resolved and if any issues have come up.

Once the data has been verified to be clean, it's saved as a permanent SAS® dataset in the study's analysis dataset folder so that programming work to create the project deliverables (e.g. tables, listings, and figures) can begin. It's common for PE data to arrive in smaller batches of patients; the DQR process repeats for each delivery until they can all be appended into a single analysis dataset and the reports rerun for a final review by the epidemiologist.

DATA REVIEW CATEGORIES

After defining the general workflow for the DQR process, we created a listing of Data Review Category (DRC) items to define what critical components and data checks were needed when applying the DQR process to project work.

Through cross-functional collaboration between various statistical programmers and epidemiologists, these categories were created and refined into four levels including:

1. Labeling and formatting of variables.
2. Identifying missing or incomplete data rows.
3. Checking key data point distributions.
4. Verifying logical consistency in categorical and quantitative variables.

The DRCs were an integral component in helping to define the programming specifications for the DQR tools and programs that still needed to be developed in SAS® and what they should check for. The time required for processing raw vendor datafiles into SAS® analysis datasets has been noticeably reduced since implementing the DQR process, since much of the previous back-and-forth ad hoc communication regarding identifying and resolving data issues has been streamlined into a much more efficient process.

THE DATA QUALITY REVIEW ISSUES LOG

A customized Data Quality Review Issues Log (DQRIL) was developed alongside the DRC for documentation, traceability, and auditing purposes. The DQRIL is a file with multiple tables that keeps track of open and resolved DQR issues at the level of each PE study. While most data issues tracked can be fixed programmatically in SAS® using existing DQR tools or with additional clarification from the epidemiologist, there are some complex data entry/collection issues that, if appropriate, can only be resolved with a new raw datafile delivery from the vendor.

The DQRIL keeps a record of key identifying information such as:

- The programmer who discovered the DQR issue.
- The date the issue was discovered.
- Which variables and datasets were potentially affected.
- Who to contact for follow-up, and the final resolution implemented.

Overall, we've found the format of the DQRIL and defined DRC items to be very helpful in reviewing issues with the epidemiologist during meeting discussions, since it provides a broad overview of outstanding issues and finer details such as which datasets, variables, or patients are affected.

STANDARD VARIABLE NAMES

The goal of standardizing variable names is to take advantage of some of the existing Clinical Data Interchange Standards Consortium (CDISC) standards and to use the available global macros and code development as a solid foundation to build our PE standards on.

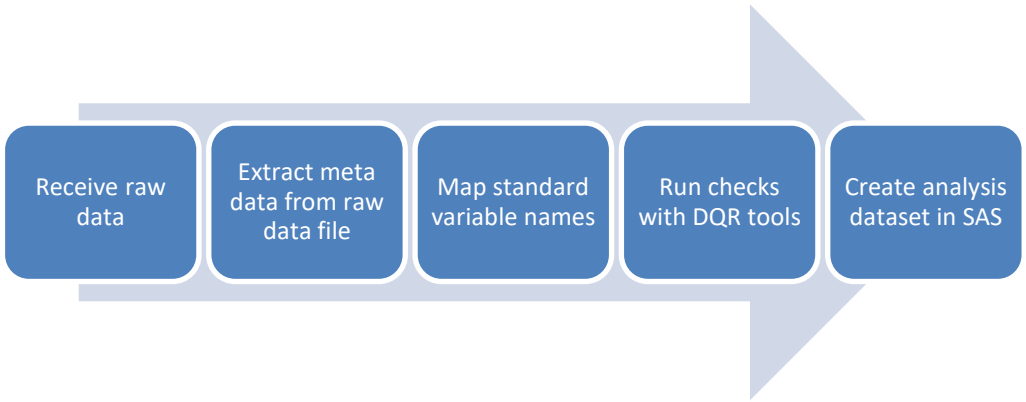
There were several challenges to overcome when trying to adopt CDISC standards and applying them to PE studies, including:

- The need for longer descriptive labels in the SAS® datasets.
- The large amount of non-CDISC defined variables present in PE data.
- The lack of a defined process and tools to convert raw variable names into standard ones.

The CDISC naming conventions were kept for all common variables; this was primarily demographic data such as SUBJECTID, GENDER, AGE, HEIGHT, WEIGHT, BMI. CDISC abbreviations such as SDT for 'start date' and EDT for 'end date' were also adopted wherever possible. These standard variable names are recorded and kept up-to-date in a reference document, so the list can continue to grow with each new study added.

The next step was development of a process and a SAS® toolset to quickly map raw variable names to standard ones. The first tool imports the raw data into SAS® using proc import procedure and exports select metadata information from proc contents into a premade define.xlsx template. This template includes a blank column for standard variable names and prefills the extracted metadata information into columns for raw variable names, variable type, length, format, and original label. From here the statistical programmer maps the raw variable names to standard using the naming conventions defined earlier and checks to make sure the parsing of the original variable labels is correctly formatted. Finally, proc import is used to read in the raw data file into SAS®, and a second macro reads in completed define.xlsx and uses it to convert the raw variable names to standard variable names along with any updates to the labels.

Depicted below is the general end-to-end process flow from receipt of raw data, standardization, data quality review and creation of analysis-ready datasets.



DATA QUALITY REVIEW TOOLS

A set of DQR macros and template code that ranges from simple to complex was developed to cover a range of functions that includes:

- Extracting detailed metadata.
- Calculating and printing out a listing of variables with missing rows.
- Identifying problematic date and other numeric variables and extracting other useful information from raw data files.

The programming specifications for these tools were based on the needs defined in the DRC. Typical DQR issues that the tools look for are:

- Extreme data outliers.
- Special or Unicode characters that do not import properly into SAS®.
- Wrong variable type once imported in SAS®.
- Various date variable issues, including multiple formats and invalid rows
- Logical checks for chronological order between date or categorical variables.

A customizable summary statistics macro provides useful information for both pre-DQR and post-DQR checks regarding the count, missing, percent missing, min, max, percentiles, and interquartile range for a broad overview of the available data. For more complex or project specific issues, there are several open code job-aid items with dummy data that provide examples on how to handle ad-hoc customer review requests.

To facilitate new programmer onboarding for PE studies and to have a defined set of instructions in place, a comprehensive execution resource was written to cover the entire DQR process and how to handle incoming raw data files. The guidance consists of mandatory and highly recommended data and consistency checks, which DQR tools to use for these checks, which tools can fix common problems, how to properly document issues in the DQRIL, and how to ensure that the final analysis dataset is truly reflecting the study raw data before analysis programming begins.

PHARMACO-EPIDEMIOLOGY STANDARDS MACROS

Through a collaborative effort between statistical programmers and epidemiologists, a comprehensive TFL package was defined and standard macros developed to create them. An initial set of TFLs was pulled and consolidated from legacy PE studies, where it was found that:

- While there were many highly varying tables, much of the statistical and numerical information presented within was the same at the level of each therapeutic area.
- There were also many tables that displayed the same information with the data presented in row (horizontal) format in one legacy table or column (vertical) format in another.

A set of standard mockups was developed to cover all of the information and statistics required and to remove redundancies caused by cosmetics or other small factors. Even though many of the TFLs were unique to PE, there was some crossover where existing global standard macros could be used, such as the one for a customizable Kaplan-Meier curves for survival analysis. Wherever possible, we tried to apply the same methodology and calculations as the global macros to preserve the sense of familiarity of use. Each of the standard mockups has built-in flexibility for rows and columns to display a wide variety of statistics, so that the customer can choose the most relevant ones for their study needs.

All PE Standards macros follow the modular programming approach for development; this ensures that the process for making future updates is more streamlined and that code can be readily reused and adapted for other needs. The final package has about 25% of the total number of TFLs as the legacy files reviewed but still covers all the required statistics and output.

CONCLUSION

Having a core standardization process in place has been greatly beneficial for reducing the time it takes to identify and resolve data quality issues with real world data; this also ensures faster and higher quality analysis deliverables for customers. The exact performance metrics are still being recorded and not yet available but based on programmer and customer feedback, the entire process has become much more streamlined and less resource intensive

Since the initial implementation of the DQR process and the PE standard macros, we seen the following improvements:

- Having the DQR tools and the accompanying execution resource has been a great help in getting programmers new to PE studies started on their project work.
- By taking advantage of the modular program approach for our standard macros, it's become much easier to update TFLs for any cosmetic changes that the customer may request after review.
- The DQR process has eliminated many of the problems encountered by PE statistical programmers during the program development phase, since key data issues are discovered and addressed well before deliverable programming begins.

We are continuing to improve the DQR process, tools, and standard macros based on programmer feedback and new customer needs. Most importantly, through this large collaborative effort towards standardization with our peers, we have strengthened the working relationship between statistical programmers, epidemiologists, and external data vendors involved. RWE will continue to expand and those of us working in the area have a unique opportunity to help guide its path towards even greater acceptance within the scientific community by implementing similar standardization initiatives.

REFERENCES

1. US Food & Drug Administration website, 2019, https://www.fda.gov/scienceresearch/specialtopics/realworldevidence/default.htm.

ACKNOWLEDGMENTS

The authors would like to thank their epidemiology colleagues at Merck & Co., Inc., Upper Gwynedd, PA, USA, for their partnership in support in this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bo Zheng
Merck & Co. Inc.
351 N Sumneytown Pike
North Wales, PA 19454, USA
bo.zheng1@merck.com

Xingshu Zhu
Merck & Co. Inc.
351 N Sumneytown Pike
North Wales, PA 19454, USA
xingshu_zhu@merck.com