Paper RE06

Leveraging Big Data in Real-world Evidence Generation: Data Engineering Challenges and Possibilities

Samiul Haque, SAS Institute, Cary, NC, USA
Robert Collins, SAS Institute, Cary, NC, USA

ABSTRACT

Medical claims, electronic health records (EHR), and electronic medical records (EMR) are major sources of real-world data (RWD). RWD can improve the efficiency of clinical trials by providing means for hypotheses testing, drug development tools identification, trial feasibility assessment, and synthetic control arms creation. In addition to clinical use cases, RWD can be leveraged for post market analysis, manufacturing quality assurance and supply chain optimization. RWD datasets are transactional data and can provide longitudinal views of patients' histories. However, appropriate data transformation, cleaning, and processing are required to ensure that users can analyze the data appropriately. Thus, the success of RWD-based studies requires close collaboration between data engineers, data scientists, and statistical programmers. In this technical presentation, we will present lessons learned from analyzing transactional claims data for real-world evidence (RWE) generation.

INTRODUCTION

The use of real-world data (RWD) to support regulatory submissions has been gaining ground, especially with the regulators now weighing in on the topic. Historically, RWD has primarily been used to support marketing safety surveillance efforts post-approval, in the commercial space and to demonstrate value. needs. RWD has also been widely used to identify potential markets with unmet need. RWD is also used in trial feasibility, trial planning and guiding site selection and patient profiles for trials. More recently, RWD is being used for external or synthetic control arms and identifying and supporting new indications or label extensions.

Real-world data differs in many ways from the data collected from randomized clinical trials. The most obvious difference is the size of the data. Whereas a clinical trial may have a few thousand patients followed for a handful of years, a real-world data source can include data for millions of patients with a decade or more of data. If the RWD source includes prescriptions, labs and observations, the number of records explodes. Another difference is that RWD is collected during the routine course of care so there are no well-defined follow-up intervals with standard measures collected at each time point like one typically sees with clinical trials. And, finally, when there is a discrepancy with a data value in a clinical trial, the collected data can be referenced against the original case report forms or medical records for clarification. Since RWD are usually anonymized, it is impossible to find the corresponding source documentation and, even if it were possible, the RWD is from the source documentation.

Utilizing RWD requires careful collaboration between the entire project team to select a suitable real-world data source; to ensure the data quality and fit for purpose; to understand the condition being studied and its course and treatment in routine care; and to apply the correct analytic techniques.

METHODS

ELIMINATE UNNEEDED INFORMATION

Regardless of the tools being applied to work with RWD, it is important to apply good practices such as only keeping the rows and columns of information needed for analysis. In SAS jobs, this is often done via a `KEEP` or `DROP` statement in a DATA Step or by specifying columns in a SQL `SELECT` statement. This one tip can significantly reduce network traffic and other I/O overhead as well as memory use, computation time, temporary workspace needs and the size of output datasets.

In order to determine what columns are required, it will be necessary to include clinicians in the discussion as well as review any data documentation. Restricting data by date or other factors can also significantly reduce the information being handled.

DATA TYPES

Different database systems may treat VARCHARs differently and since SAS trusts the metadata it sees, you must be careful when pulling in VARCHAR data. SAS will often automatically create CHAR variables with a length of 32,767.

This can be addressed with the DEFAULT_VARCHAR_SIZE option. You can also cast the length of variables by using explicit pass-through queries.

You may also encounter issues with Unicode or other character encodings so always verify your results.

DATA REVIEW

Traditional approaches to data profiling may be difficult due to the data sizes. SQL queries can be used to assess data ranges, outliers, counts and other summaries. In lieu of PROC FREQ or a DATA step for data review or for testing the results of SQL joins, it can be useful to use a PROC SQL subquery to report summaries such as the number of rows per patient:

```
select distinct num_rows, count(*) as freq
  from (select distinct patient_id,
                        count(*) as num_rows
          from join_result)
  group by num_rows
  order by num_rows;
```

A similar approach can be used to identify and obtain counts of the most common diagnoses, procedures and treatments.

DEALING WITH MISSING OR ERRONEOUS VALUES

As referenced previously, missing or aberrant values are likely to be found in any data source. It is necessary for the project team to decide how to address each situation knowing that each choice has its own benefits and caveats. It may be decided to omit the value or simply set it to missing. Other choices could include omitting the observation or omitting all records for the patient in question. Imputing a value using any of various established methods is another approach. In all cases, the choices made for each instance and rationale for the decisions should be documented.

SAMPLING DATA

Working with samples of the data can be very helpful for initial review of the data and for testing code or exploratory analysis. A more efficient method of sampling data than using PROC SURVEYSELECT or similar approaches is to simply use the last one or two characters of the patient identifier. Identifiers in anonymized data can usually be assumed to be sequentially assigned, i.e., random. For example, if the patient identifiers contain only numeric characters, choosing a single last digit such as “4” will yield a reproducible 10% sample without having to perform joins across different tables to find the correct subjects. If you need smaller samples, you can select one or more 2-character strings from the right-hand side of the identifiers where each two-character string will select 1% of the data.

MARKET DEFINITION TABLES

Another way to save time across all projects is to create “market definition tables” which are essentially tables of diagnoses, procedures or medications associated with a specific condition. These tables can then be used to identify patients or records meeting those criteria without having to look up and create a list of codes each time.

SQL PASS-THROUGH

Understanding and using SAS’ SQL pass-through capabilities can significantly improve performance of jobs and simplify some activities. When a query is “passed-through” to the source database system, that query is executed by that source system so it can take advantage of any performance enhancements native to that system. Another benefit of a pass-through query is that it reduces data movement because only the results are transferred to SAS instead of all of the records. For explicit pass-through, you will typically need to first establish a connection to the database system and then be certain to express the query in the syntax of that environment. The example below illustrates create a connection to an Oracle server that uses Oracle’s to_date function to create a date from a string:

```
connect to oracle as myconn (user=smith password=secret path='myoracleserver');
  select *
    from connection to myconn
         (select patient_id, drug_id, servc_date
            from t_rx
            where servc_date < to_date('31-Dec-2022', 'DD-MON-YY') );
disconnect from myconn;
quit;
```

TOOLS FOR RWD ANALYSIS

There are a number of tools that can help perform analysis of real-world data. One simple approach is the use of visualization tools to explore your data distributions and check for outliers. Mapping tools can also help you identify bias or distribution of data geographically. In addition, OHDSI ATLAS and other commercial platforms, including SAS, provide a number of different tools and utilities to enable users to more quickly specify the patient populations they are interested in studying.

USE CASES

ADVERSE EVENT SIGNAL DETECTION

One of the world's largest medical device manufacturers is leveraging SAS Viya to detect and report adverse signals collected from their medical devices. Applications on SAS Viya mine terabytes of real-world data from data marts and escalates important device signals to proper stakeholders. In this case, SAS Viya enabled them to scale and productionize signal detection and escalation applications so that anyone in their organization can use these apps to escalate issues and generate reports without assistance from programmers or analysts. By leveraging SAS' CAS and Data Connectors they have increased the performance of their ETL and data preparation programs. The integrated deployment of these applications reduced their costs, number of tools, and time to discovery while impacting patients lives in a positive way.

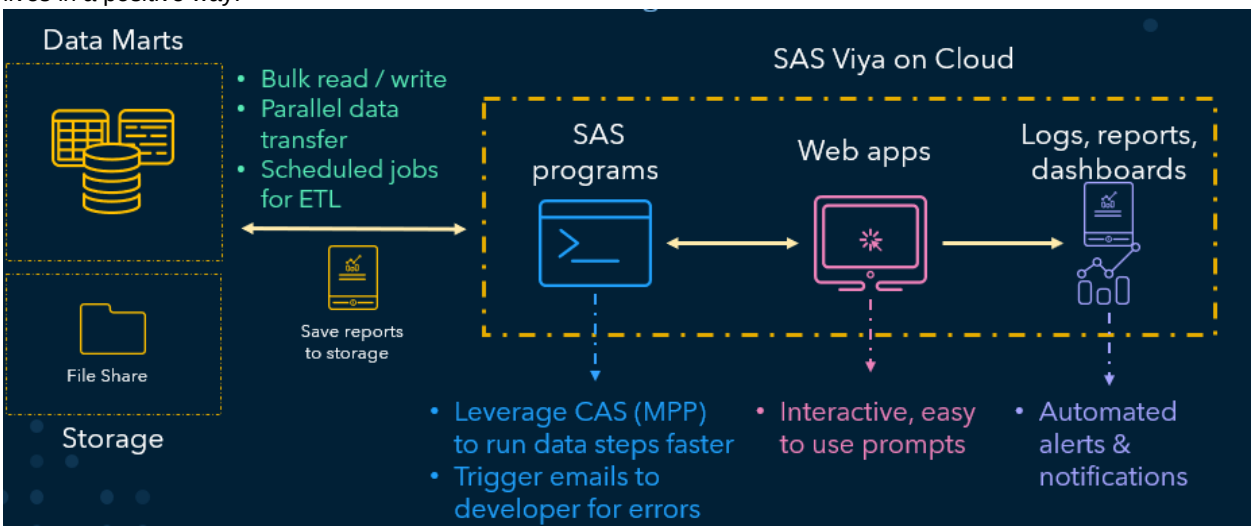

Figure 1. Adverse Signal Detection and Escalation . Application

SELF-SERVICE COHORT ANALYSIS

A leading health care data provider has built a self-service dashboard on SAS Viya to understand medication adherence (persistency analysis). This dashboard allows analysts to:

- Select a market of interest (e.g., NSCLC, diabetes, MCL) and build a corresponding cohort of patients;
- Modify analysis parameters (e.g., list of drugs, grace period);
- Run templated analytics; and
- Dynamically extract data from the patient data mart.

This enabled this data provider to:

- Respond faster to customer queries;
- Streamline prescription persistency analysis processes;
- Reduce the number of tools to be used and to support; and
- Avoids creating multiple copies of these large datasets.

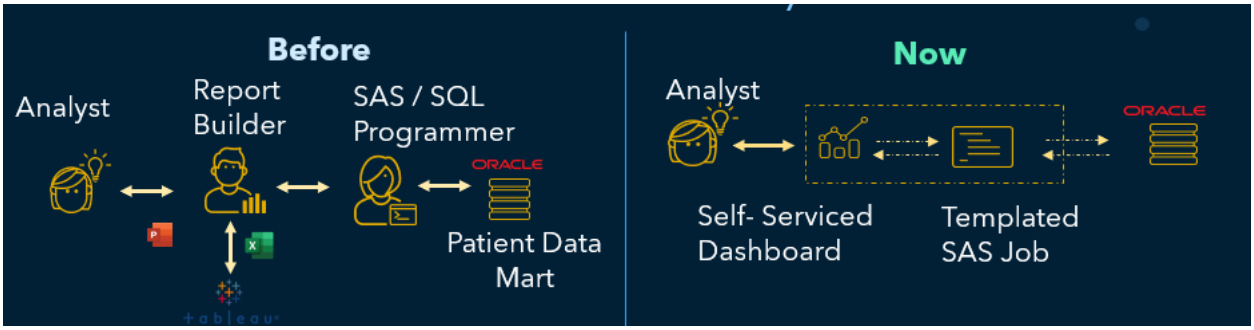

Figure 2. Self-Service Cohort Analysis Dashboard

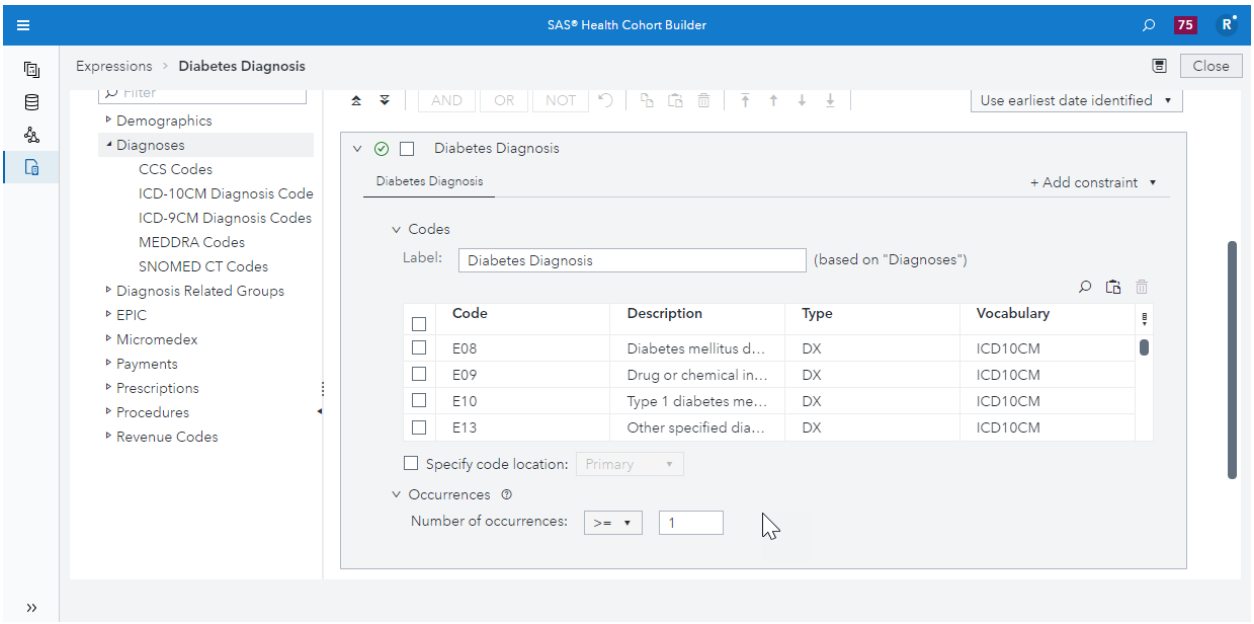

Figure 3. In this screen of a graphical cohort builder tool, the user has selected the codes to be used to identify a diagnosis of diabetes and added them to the cohort search criteria. From here, the user can continue to add or edit criteria or move to the next step of the tool.

TRIAL FEASIBILITY

Leveraging real-world data can enable researchers to explore actual populations and their behavior and outcomes in the course of routine care. They can also test hypotheses rapidly to determine clinical relevance along with care and treatment pathways of the desired cohort. Researchers can even model a clinical trial by applying predictive analytics. Machine learning can remove the need for researchers to manually test hypotheses and shorten the time from hypothesis generation to optimized trial design. The expanded availability of RWD sources and encouragement by regulators suggest that using RWD to optimize eligibility criteria and recruitment will become the “new normal” across the clinical trials enterprise.

EXTERNAL CONTROL ARMS

External or synthetic control arms have been more common in medical device and rare diseases because of the ethical issues and the small sample sizes. However, regulatory agencies including the US FDA have been increasingly providing guidance and encouragement for this use of RWE in clinical trials. However, this process has many issues that must be carefully considered and addressed. One common difference is without the randomization of a typical randomized clinical trial, methods must be applied to ensure a fair comparison between study arms. This is often addressed with propensity score matching where propensity scores or the “likelihood to be treated” are calculated - typically using logistic regression across selected covariates – and then controls from the real-world data are matched to cases from the treatment arm based on similarity of the propensity scores. These matches can be 1:1 or many-to-1. A critical step in this process is to assess the quality of the matching applied to ensure the control arm is not biased. Other issues with this approach include deciding whether to do retrospective or prospective data for the control arm; how to deal with the lack of defined follow-up times and measurements in the control arm vs. the treatment arm; and many other factors.

POPULATION HEALTH

RWD can be used to understand and analyze issues affecting population health such as patterns of disease distribution, progression of disease, economic and quality of life impacts of disease and many others. Another aspect of this type of work is to analyze the impact of disease management or care management on patients to determine the effectiveness of the program. In these types of analyses, it is often discovered that patients' weighting of important factors in their care and disease or often different than proposed by clinicians and researchers.

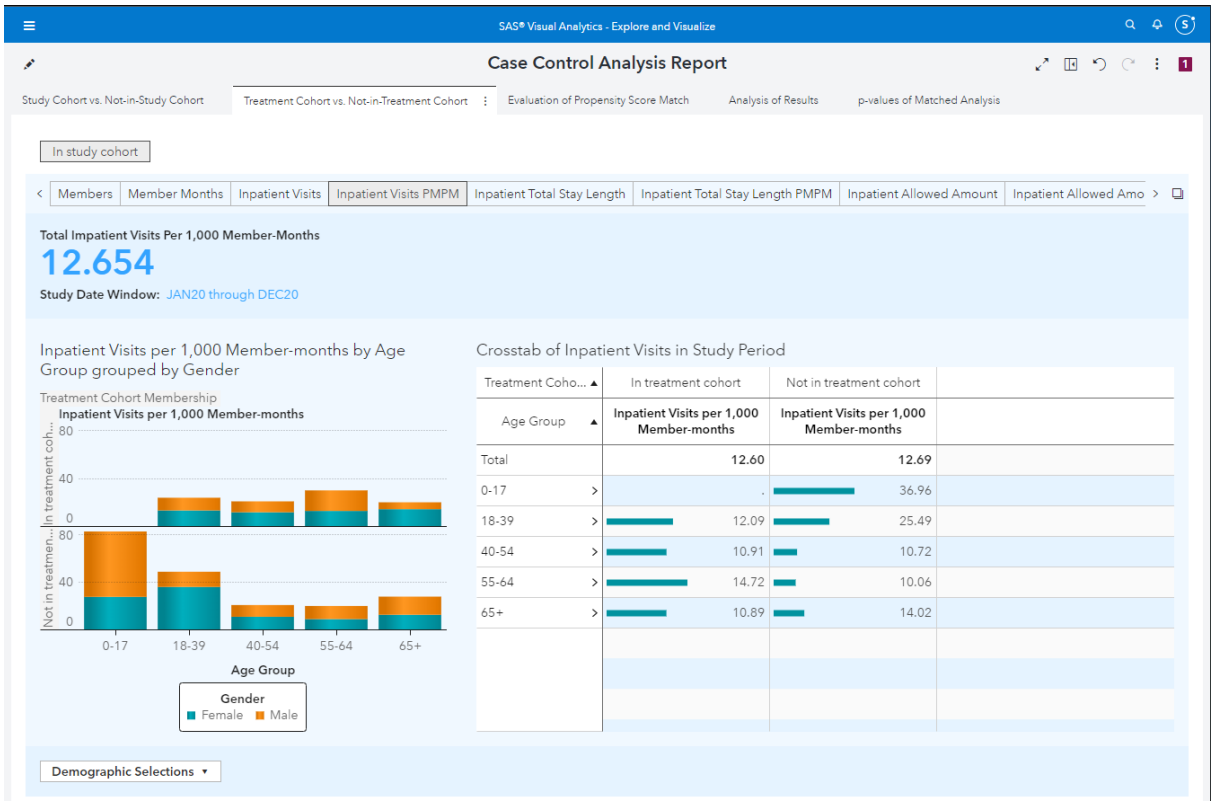

Figure 4. This image is from a report produced on a hypothetical disease management program with the tab shown reporting the differences in inpatient visit rates per member per month broken out by age group.

OTHER USE CASES

Many other use cases for these types of analysis built on real-world exist and can be accelerated by use of tools designed for working with these large datasets. In addition, special attention has to be paid to ensure the correct analytical methods are applied and potential confounding issues from the data sources are understood and minimized. Other examples might include areas such as:

- Pharmacovigilance
- Other patient outcomes including comparative effectiveness of treatments (including quality of life and economic factors at both the individual and population levels);
- Identifying early or targeted interventions that may improve outcomes;
- Patient journey including medication persistence, compliance and switching and their impacts on outcomes;
- Expand or new indications for existing treatments or drug repurposing;
- Commercial and market support including establishing the value of a treatment and gaining formulary access or understanding access issues, prescribing patterns and supply chain issues.

CONCLUSION

There are many opportunities for the use of RWD beyond the clinical domain and this effort may require some different techniques than those used in clinical data analysis. With a thoughtful analysis of real-world data, you can impact other parts of your organization and improve patients live in a positive way.

REFERENCES:

US Food & Drug Administration. (December 2018). *Drug Framework for FDA's Real-World Evidence Program* https://www.fda.gov/media/120060/download

CONTACT INFORMATION

Samiul Haque
SAS Institute
SAS Campus Drive
Cary, NC 277513
+1 919.531.2797
Samiul.Haque@sas.com

Robert Collins
SAS Institute
SAS Campus Drive
Cary, NC 277513
+1 919.531.0051
Robert.Collins@sas.com