PharmaSUG 2026 - Paper RW-384

Assessing Quality of Real-World Data Sources

S. Robert Collins, SAS

ABSTRACT

The FDA continues to update guidance on the use of real-world data (RWD) and evidence (RWE) for drug and biologic products and medical devices. While regulators are increasingly open to the use of RWD sources that have flourished as a result of technical advancements and requirements such as medical data interoperability, sponsors and regulators believe there are still significant opportunities for the use of RWD as long as the quality and provenance of the sources are demonstrable.

In this session, we present a discussion applicable to anyone interested in the use of RWE of “first-person” and “third-person views” of data by the generating institutions and the consumers of pooled data and establish that source-based quality methodologies have the greatest potential for improved data. We next review the factors that give rise to problematic data at the source. We then look at practical approaches for a third-party data consumer to assess and document data quality. These methods present and expand on traditional methods of exploratory data analysis and consider the use of machine learning and artificial intelligence to provide new approaches to assessing data quality.

This paper first reviews the evolution of real-world data sources, then examines factors that affect data quality at the point of collection. We next discuss practical methods for assessing and mitigating data quality issues from the perspective of third-party data consumers, with particular emphasis on exploratory and programmatic validation techniques.

INTRODUCTION

The nature of real-world data is undergoing significant changes. The primary source of RWD used to be administrative data such as Medicare and insurance claims which were collected for reimbursement purposes and contained limited clinical detail. Registries were also used but they tended to be limited in the population size, restricted to patients sharing a specific condition or treatment, and difficult to combine if you needed a larger sample size.

More recently, the digitalization of healthcare has been driven by factors such as adoption of electronic health records or electronic medical records followed by growing standards and interoperability requirements and regulatory efforts including:

- American Recovery and Reinvestment Act of 2009 (ARRA)\
- HIPAA
- HITECH
- 21st Century Cures Act
- ONC HTI rules
- TEFCA
- CMS interoperability rules and
- Health Information Exchanges (HIEs).

However, the movement toward widely-accessible clinical data has been slowed by implementation issues and privacy concerns.

We will use the terms “first-person” to describe data collected by an organization during the delivery of direct patient care and “third-person” to describe data that has been aggregated by an organization not directly responsible for patient care. These 3rd-person sources include a variety of vendors that provide access to de-identified patient records.

FACTORS IMPACTING QUALITY OF SOURCE DATA

First-person and third-party perspectives differ in terms of incentives and visibility into the data-generation process which has implications for how data quality should be assessed and documented. First-person

data comes directly from the collecting institution and may represent some opportunity to discuss issues and potentially improve subsequent data transfers. The primary concern of a 1st-person data generator is patient care and not downstream uses of the information. Third-person data is more commonly used across the industry and represents data aggregated across multiple providers which provides a richer population of patients. These 3rd party data vendors often apply their own data quality checks and edits. These vendors are focused on assembling a representation population that can be marketed to researchers. It is critical to discuss with the vendor any cleaning they may have performed. Most 3rd party data providers also work to ensure a consistent anonymous patient identifier across all of their sources which means you may have access to records from multiple providers which may not be available with 1st person data.

A number of factors at the point of data collection can impact the quality of RWD. While electronic health records and electronic medical records (EHRs) have improved the situation, there are still areas that impact the quality of the data. The simplest issue is human error such as transposing numbers during entry or inconsistent units. Where structured data are collected, there can be inter- and intra-observer variations in collection. One observer may skip some items in some circumstances but not others. Another observer may give primary focus to a different set of values or enter data more frequently. For elements recorded using free text instead of coded values, it can be extremely difficult to standardize entries. The standardization of data collection is typically secondary to patient care activities whereas RCTs focus on accurate data collection as an adjunct to clinical care.

Often, data elements are simply missing with no indication for the reason. Unfortunately, we know that missing data are often not “missing at random” and could represent some process issues. The underlying causes of missing values can potentially create bias in the data.

As mentioned previously, administrative claims data typically do not capture rich clinical data. The recorded number of diagnoses or procedures may be capped and have probably been adjusted to optimize reimbursement. In addition, claims data focus on diagnoses and procedures but do not capture observations or results. Medical records themselves can generate data quality issues. How is older information updated when newer information comes along? Keep in mind that any data collected is a “point-in-time” and may differ from the patient’s current state. “White coat syndrome” where anxiety or stress causes a patient’s blood pressure to rise in a medical setting but returns to baseline when at home is one common, simple example of this. In addition, there is often no information in the medical record indicating medication compliance so just because a patient has been prescribed a blood pressure medication doesn’t mean they are on it during their visit. Conversely, they may have started taking a medication they had been ignoring because they know they have an upcoming appointment.

Even with an EHR, how are older data migrated into the system? Did this process leave gaps? Note that each institution typically customizes the EHR to their preferences – the same vendor’s product in different institutions will likely be different. In addition, different services within an institution may have different EHR implementations or vendors. All of these can complicate harmonization efforts and introduce another source of system variation that is difficult to detect in pooled data.

DATA QUALITY CONSIDERATIONS: RCTS VERSUS RWD

Randomized controlled trials (RCTs) have well-established methods from study design through regulatory submission to address the quality and validity of the data and analytical results. Unfortunately, those methods cannot simply be transferred to real-world data sources for a variety of reasons. That is not to say that there is no guidance for working with RWD to generate evidence. RWD is established for post-marketing analyses. RWD is also used to address practical and ethical considerations such as medical device approvals, rare diseases and withholding potentially-lifesaving treatment from a group. Ideally, RWE research should apply the same approach to careful planning, documentation and transparency to ensure results are accurate and reproducible.

RWD requires different analytic methods and considerations due to the lack of randomization, especially when investigating causality. The most common approach to address this is through the use of propensity score matching (PSM). PSM allows the creation of matched samples of patients who did or did not receive a treatment. These samples are created based on patient characteristics and are intended to

balance selection and preference biases. Logistic regression is the most common method of calculating propensity scores but other methods can be applied. Rubin, Cochran, and others developed the methods in the 70's and 80's specifically for observational studies. Note that these PSM methods only address *observed* confounding and ISPOR and other groups are working on ways to identify and address unobserved bias and confounding.

One primary difference between RCT data and RWD is that RWD are typically anonymized which makes tracing back to the source impossible. Even if you could trace back to the source, the RWD contains what was originally recorded which would make any corrections based on interpretation which introduces yet another source of bias.

EVALUATING SOURCE DATA QUALITY

DATA RECEIPT

The first steps for assessing data quality are essentially "trust but verify." Begin with comparing the data transfer documentation's row counts, column names and data types with the received transfers and with the data once it has been loaded into the local environment for analysis. Some transfer methods can be trickier to import than others but these basic checks are always a worthwhile investment. These approaches can help you identify issues with the data extraction and transfer and, more likely, with your processes of loading the data into your environment.

Another critical step is looking at counts by keys in individual tables. For example, if a table is only supposed to have one row per patient ID, check to ensure that is the case. If every patient is supposed to be represented in the demographic table, check to ensure none of the other tables have patients that are not in the demographic table.

Once the data are available in the environment for analysis, you can perform more traditional exploratory data analysis methods including simple frequencies and other approaches for identifying outliers. This step may identify additional issues with the data transfer and documentation – you may find values that do not match the data dictionary. When this happens, you need to decide and document your actions. The solution is often as simple as updating the data dictionary to match reality.

There are many sources that can help you develop your EDA methods. Keep in mind that sometimes you may desire to check temporal relationships in addition to univariate statistics. These checks can identify issues such as procedures performed before related diagnoses which may indicate data censoring.

Nested SQL queries can be used to easily perform many data checks either within or between tables. One example of this would be to examine the distribution of records each patient has in a certain table – nested queries can easily generate a list reporting observed patient-level record counts in descending order with the number of patients having each of those record counts.

While you can reduce the time and effort to perform EDA by parameterizing and reusing code and using data visualizations in addition to tables, recently interest has grown in automating exploratory data analysis. Many applications and tools now incorporate these features. There is a growing number of open source packages that provide these capabilities and the SAS® Viya® platform offers its Information Catalog and other methods to accelerate these activities. As with every other approach, it is important to understand what is and is not covered by these tools to ensure that you provide full coverage in your data quality checks.

SUBSEQUENT TRANSFERS

It is recommended to keep the code you use for verifying the received data and for EDA so that you can rerun the same code when new data are received. You'll also need to know if the new data includes corrections to existing records and, if so, how to apply those. Recheck everything, especially any previously identified deviations from the documentation. If data types or coded values have changed, you may need to adjust your existing code. It is recommended to use code versioning and change logs to document adjustments. It is also beneficial to document changes across refreshes – Are demographics changing over time? Are comorbidities for the patients of interest changing? The introduction of new treatments and therapies may also impact analyses.

ANALYTIC DATA PREPARATION

As you write code to create analytic data, it is important that you maintain your internal quality control processes. Version your code and track changes. To improve validation and reproducibility, one simple approach when you want to separate out records for analysis, you may choose to create separate output datasets for the selected records and the rejected records. This enables easier review of the inclusion and exclusion logic. A simple way to do this in a SAS DATA step is to include two datasets on the DATA line and conditionally write every record to the appropriate destination. You may choose to print a subset of the deleted patients to the log emphasizing the discerning variables and then delete the dataset knowing the record counts are in the log.

When using SQL for joins, it can sometimes be difficult to tell if you really got what you expected. Verifying the results of joins is critical for controlling the quality of your analytic data. If you expected one record for patient in the results, check to ensure you really have one record for patient.

METHODS FOR ADDRESSING SOURCE DATA QUALITY ISSUES

There are many methods for addressing suspicious data. Preferably, these situations will be captured during your data quality assessments and then decisions on handling the issues can be determined by all of the stakeholders at analysis time. When data issues are identified, there are several common response strategies:

- Eliminating the patient from the analyses;
- Eliminating the record from the analyses;
- Eliminating the value from the analyses;
- Imputing a value for the analyses;

There are tradeoffs for each of these decisions or other methods that may be applied but in all cases, your decisions should be documented and reported in your results.

If possible, you should compare what you see in your data against either other data sources or published information. Each data source may have its own biases. For instance, private insurer data will typically represent younger, wealthier, more-educated patients than Medicare. There are also differences due to geographic representation. These can include different ethnic compositions, different diets or access to care. These differences may be expected or even desirable if one source better addresses the research question.

CONCLUSION

There are many considerations for ensuring the quality of real-world data – from source to analysis – and it is the responsibility of the team using the data to ensure everything is done to ensure the data are as clean and accurate as possible. Unfortunately, RWD present some issues not seen with RCT data. For this reason, documentation, transparency and reproducibility are critical. By treating data quality assessment as an explicit, documented analytic task rather than an assumption, researchers can improve the credibility and regulatory utility of real-world evidence.

CONTACT INFORMATION

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

S. Robert Collins
SAS Institute
Robert.Collins@sas.com
https://www.linkedin.com/in/srobertcollins/

AI Disclaimer: Microsoft Copilot was used to proofread the content and suggest changes.