Paper RE05

Keeping It Real: A Tiered Approach to Real-World Data Quality by Leveraging Synthetic Data

Lisa Murch, SAS Institute, London, United Kingdom
William Yuanhao Kuan, SAS Institute, Tokyo, Japan

ABSTRACT

Data quality remains one of the most significant barriers to robust real-world evidence (RWE), particularly in Europe, where fragmented healthcare systems, local coding standards and data formats, and GDPR restrictions complicate access. We outline a tiered approach to improving real-world data (RWD) quality with SAS Viya, a data and AI platform. We begin with data profiling and rule-based checks to identify common data quality issues such as missingness, temporal inconsistency, and non-standard codes across EHR, claims, and registries. Next, we explore the potential of synthetic data for addressing the challenges of data quality, and discuss how synthetic data can be used to fill the data gaps, mitigate bias and enhance privacy compliance. Finally, we explore the challenges of integrating proprietary and open source tools while maintaining compliance with regulatory standards and the goals of European RWD networks such as DARWIN EU.

IMPORTANCE OF DATA QUALITY OF RWD

In the new era of evidence-based decision making, RWD provides an additional layer of insights on patient health information beyond the randomized clinical trial. RWE captures the diversity of daily clinical practices and enhances the representativeness of clinical evidence alongside a well-controlled clinical trial.

However, these advantages are critically dependent on the quality of the underlying RWD. Inaccurate, incomplete or biased RWD can mislead researchers and decision makers and as a result, fail to deliver value to patients. Data quality is central to the meaningful use of RWD and generation of RWE in the pharmaceutical and life science industry. Poor quality of RWD will affect regulatory decisions, drug safety and effectiveness evaluation, and ultimately patient outcomes. High-quality and fit-for-purpose RWD ensures that evidence derived from it is reliable, accurate and reflects the real-world clinical situation.

CURRENT CHALLENGES IN DATA QUALITY OF RWD

While pharmaceutical and life science companies are incorporating RWD into product development strategy, they face multiple challenges. One obstacle is the lack of standardized data collection practices. Variability in format, methods and terminologies across providers makes it difficult to compare and analyze the data in a comprehensive way. It is not uncommon for data to be incomplete or biased. It is challenging to clean, manage and organize this fragmented data in a way that could capture the comprehensive patient journey.

Beyond those limitations inherent to RWD, life sciences companies also face challenges in methodology and process when generating RWE from RWD. For example, many organizations lack harmonized frameworks to control the quality of RWD before starting analysis. Without frameworks developed by experienced data engineers, transparent and reproducible processes are challenging.

GUIDELINES ON DATA QUALITY OF RWD

Regulatory agencies worldwide acknowledge that data quality is a critical factor in ensuring the reliability of RWD. They have published several frameworks and guidelines addressing different dimensions of data quality and the common principles of handling data quality issues.

In Europe, the European Medicines Agency (EMA) has established its Data Quality Framework (DQF)[1], which serves as a structured approach for assessing the quality of RWD. The DQF outlines five key principles: 1) Extensiveness; 2) Coherence; 3) Timeliness; 4) Relevance; and 5) Reliability. Transparency in data handling is essential for its use in regulatory submissions. On the other hand, the U.S. Food and Drug Administration (FDA) takes a slightly different approach. FDA guidelines emphasize evaluating data quality from the dimensions of completeness, conformance, plausibility and integrity[2]. A thorough documentation of data collection and management practices as well as quality assurance and quality control plans are required for submission purposes.

APPROACHES TO ADDRESS DATA QUALITY ISSUES

To address the data quality issues systematically, we recommend that an organization takes multiple-pronged strategies. A strong foundation begins with company-wide data quality frameworks, policies, and governance structures that embed quality checks across the data lifecycle. Once the frameworks are established, it is crucial to sustain data quality through continuous monitoring, regular audits, robust user feedback collection and transparent documentation of the data transformation process. To avoid data silos, a centralized data procurement team should help streamline access to data assets across departments. Finally, technology also plays a key role in addressing data quality issues. We can identify outliers via machine learning, visualize missing data and other data quality issues by using dashboards as a control center, and capture issues in unstructured text data through natural language processing.

AUTOMATED DATA QUALITY APROACH

In the following sections, we discuss how SAS Viya, the high-performance AI & Analytics Platform, can support the development of a robust enterprise solution to strengthen the credibility of RWD and RWE, by providing the building blocks for an automated end-to-end data quality workflow. By combining automated checks, unsupervised machine learning and human-in-the-loop oversight, users can surface and investigate quality issues in large data sets and take action to remediate them. All steps are transparent and reproducible, with the option to embed Python or other open source modules as necessary, and can enable evidence-based decisions on the use of data to be traced and validated.

The approach takes several steps:

1. A strong **Data Quality Tier** to surface issues via profiling and automated quality checks, based on rules designed to ensure data is fit for purpose (i.e. domain-specific, or based on frameworks from regulators or other guidelines)
2. An **Analytic Investigational Tier** that enables visual investigation and unsupervised and supervised machine learning methods to surface anomalies and other trends that might signal data quality issues for the review of subject matter experts.
3. An **Action Tier**, to investigate suspicious records and where necessary, address bias or data gaps via the creation of synthetic data to ensure upstream analytics are as robust as possible.

1. THE DATA QUALITY TIER

SAS Viya offers out-of-the-box features for data profiling

AN EXAMPLE OF AN AUTOMATED RULE-BASED DATA QUALITY WORKFLOW FOR THE DATA QUALITY LAYER

In this paper, we describe the process of building the automated quality check component for the data quality layer.

DATA SOURCE

This study used artificial data from the Simulacrum, a synthetic dataset developed by Health Data Insight CiC and derived from anonymous cancer data provided by the National Disease Registration Service, NHS England[3]. We modified some values, to introduce some known data quality issues into the data set, in order demonstrate how an automated workflow can be used to assess the quality of the datasets and identify the records with issues.

DEFINE THE RULES OF DATA QUALITY CHECK

The first step is to define the rules of data quality check. We referenced the EMA's Data Quality Framework[1] to set up our rules based on the dimensions of Reliability, Extensiveness, Coherence and Relevance. We did not address Timeliness in this paper.

Some example rules are in Table 1.

Table 1 Example of Data Quality Checking Rules

Dimension	Description	Example
Reliability	-Data values and distributions agree with internal measurements or local knowledge. -Logical constraints between values agree with common knowledge. -Observed or derived values conform to expected temporal properties.	-Height and weight are positive values. -The patient's sex agrees with sex-specific contexts (e.g., pregnancy) -Discharge date/time happens after admission date/time
Extensiveness	-Missing required values. -Coverage of a population.	-Sex (or Gender, if used) should not be missing
Coherence	-Unique (key) data values are not duplicated.	- A patient ID number is only assigned to a single patient.

	-Data values conform to allowable values or ranges. - Data values conform to the representational constraints based on external standards. -The precision of values is fitting a target standard	- Sex only has values “M”, “F” or “U”. - Diagnosis code conforms to ICD-10 standards. - Two decimal digits are used in continuous variable.
Relevance	- The number of variables available in a given dataset matches the number of required variables	- All the necessary variables based on the research questions are available.

In this demonstration, we used the following data quality rules as an example to check against the datasets:

- No duplicate patient ID number.
- No unexpected value in “Gender” variable other than “Male” and “Female”.
- No unmapped ICD-10 code in “Cause of Death” variable when referring to ICD-10 code dictionary.
- If a patient is marked as “Dead” in the “Vital Status”, the patient must have a valid value in “Cause of Death”
- Date format in “Diagnosis Date” should be YYYY/MM/DD.
- “Age” variable should be YEAR in unit and be integral, no decimal, and within the defined range (0-130).
- Charlson Comorbidity Index score should be integral, no decimal, and within the defined range (0-37).
- “Height” variable cannot be negative number and should be reasonable range (0.4m-3m).
- Drug name in “Regimen” variable should be standardized to lower case and with no leading blank.
- Biological female cannot have prostate cancer as the “Cause of Death”.
- “Treatment Start Date” should always be less than or equal the “Date of Death”.

BUILD AND MANAGE THE RULES

We used SAS Intelligent Decisioning to build and manage data quality check rules[4]. SAS Intelligent Decisioning provides a powerful cloud-based platform for generating, managing, and monitoring business rules. This system can be used to orchestrate data quality across enterprise systems. By enabling organizations to define and automate complex rule logic, it helps validate incoming data, detect anomalies, and enforce consistency. Users can work with its intuitive GUI and integration capabilities to create and export reusable rule sets, deploy them across multiple channels, and continuously monitor performance to ensure compliance and accuracy.

First, we created the “Rule Set”. A Rule Set is a collection of logic statements that define conditions that can be used to evaluate data quality. For example, a Rule Set can check whether patient ages fall within a valid range, or whether mandatory fields like diagnosis codes are populated with a valid code. As shown in **Figure 1**, we set up a rule to check whether a record has unexpected value in “Gender” variable other than “Male” or “Female”.

Rule Sets can be versioned, tested against sample data, and deployed as reusable components across multiple data validation projects. This ensures consistency in applying the same rules to different datasets or systems. Multiple rule set are managed in the repository with the info of “Date Modified”, “Modified By”, and each published rule set is versioned for full governance and traceability (**Figure 2**).

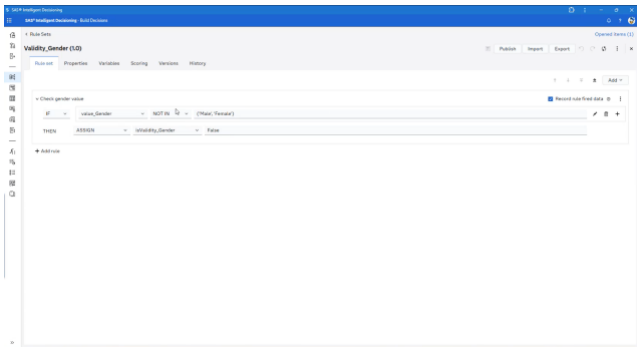

Figure 1 Rule Sets

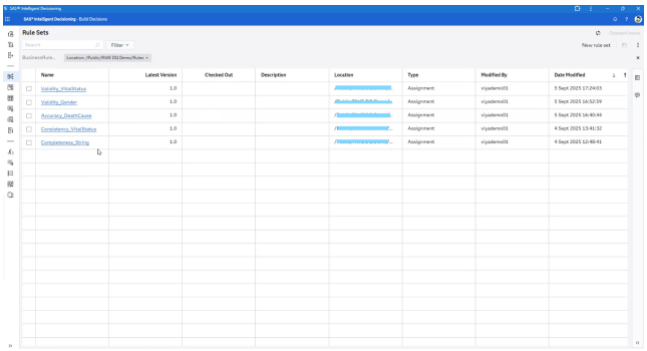

Figure 2 Repository of Rule Sets

Data quality checks often require validation against reference standards, terminology or dictionary, such as a list of valid disease codes or drug codes. A “Lookup Table” is a key-value pair structure that simplifies rule logic by externalizing reference data. Instead of hardcoding multiple IF statements, we can store reference values in a table and use lookup functions. For example, we can check whether there are records with invalid ICD-10 codes in the variable “Cause of Death”.

We created a table of acceptable values or mappings of ICD-10 codes dictionary (**Figure 3**. Not the full list). The table can be referenced inside Rule Sets, so that any incoming data is validated against the list. Maintaining Lookup Tables can prevent errors due to invalid, outdated, or inconsistent reference values.

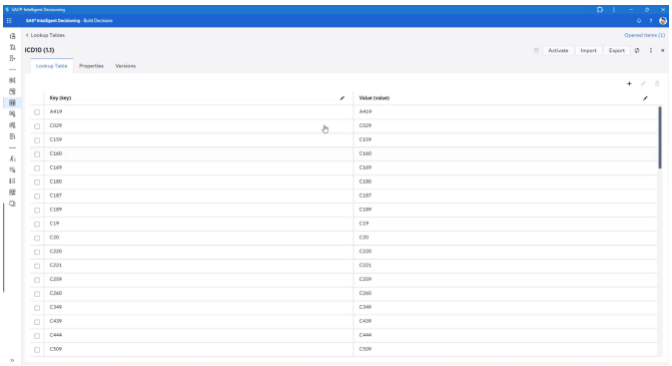

Figure 3 Lookup Tables

Once the Rule Set and Lookup Table are set up, we further combined those together into a “Decision”. A “Decision” in SAS Intelligent Decisioning is a container that orchestrates multiple components, such as- rule sets, lookup tables, analytical models, and Conditional Logic, into a single decision flow. The Decisions feature acts as the orchestration layer for executing data quality rules at scale. While Rule Sets define “what to check,” Decisions define “when and how to check.”

For example, as shown in **Figure 4**, the Decision Flow is to check if a patient is marked as “Dead” in the “Vital Status”, the patient must have a valid ICD-10 code value in “Cause of Death”. It will check against an external ICD-10 code dictionary to decide if the value is valid.

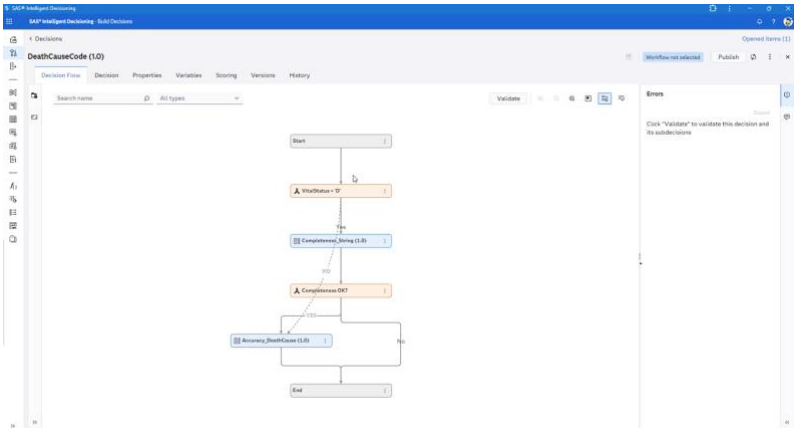

Figure 4 Decisions

DEPLOY THE RULES INTO A LARGER WORKFLOW

Once the decisioning flows are ready, we can deploy them into a even larger workflow in SAS Studio Flow to build an fully automated end-to-end data quality check workflow. This approach allows the users to include more complex data management, cleaning, and data wrangling tasks, and to export the final assessment results into a interactive data quality dashboard.

SAS Studio Flow offers a visual, low-code environment. Users can design end-to-end processes that ingest, cleanse, validate, and report on data using prebuilt tasks and custom logic[5]. These flows can incorporate decision flow from SAS Intelligent Decisioning, enabling seamless integration between decision logic and data preparation. With scheduling and orchestration features, SAS Studio Flow ensures that data quality checks run consistently and reliably, supporting scalable and auditable data governance.

For example, as shown in **Figure 5**, the original datasets were imported into SAS Studio, and taken through standard data pre-processing tasks. Next, multiple Decision nodes were executed to address multiple data quality dimensions such as validity, accuracy, completeness etc. Then, it identified all the records with possible data quality issues, combined and transformed multiple datasets with bad records, and exported those bad records into a interactive data quality dashboard for review. Although not shown in the screenshot here, the datasets with good records flagged as “No Issues” were stored separately, and the datasets with bad records were routed to a remediation process to enable the data quality issues to be recorded and addressed as appropriate. The full SAS logs are recorded and available for transparency and traceability purpose (**Figure 6**).

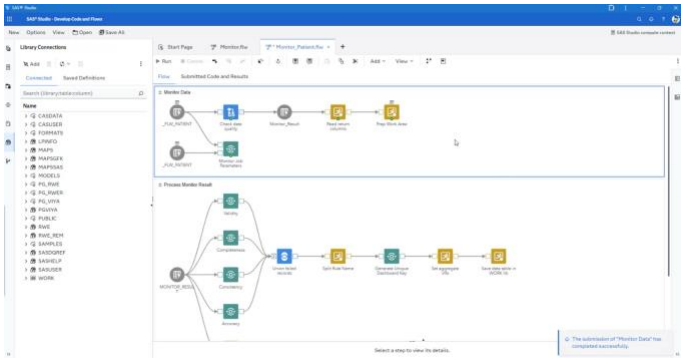

Figure 5 Fit the Decision nodes into a larger Data Quality Check Workflow

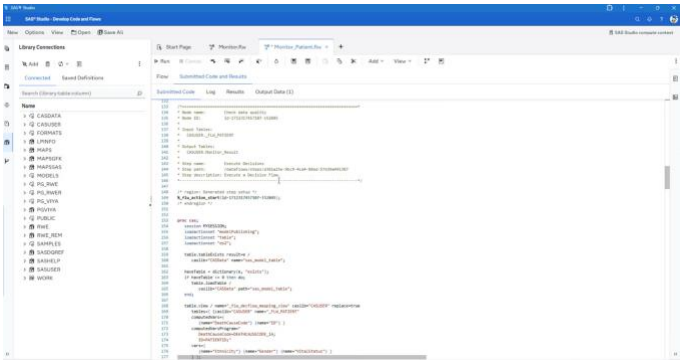

Figure 6 Full SAS logs and codes for transparency

INTERACTIVE DATA QUALITY MONITORING DASHBOARD

As the entire Flow can be scheduled to run in any time interval or trigged by an event, the data quality check results can be automatically pushed into an interactive dashboard for continuous monitoring. It is a critical step to help data engineers make results visible and actionable. An interactive dashboard provides a central, real-time view of data quality metrics and issues, turning raw validation results into insights for decision making.

As shown in Figure 7, the summary report shows the number of total records processed in the table named "TUMOUR", the number and percentage of records with issues in each variable, types of rule violations and more. When the user drills down to the issues on the "Age" variable, it shows two patient records with age info out of the predefined range (i.e. 0-130 years old). For another table named "PATIENT", one patient record was identified with Gender recorded as "Female" but with prostate cancer as the cause of death, which is a logical error (**Figure 8**).

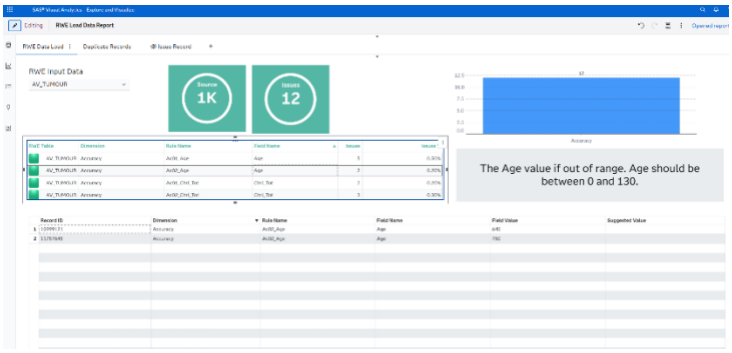

Figure 7 Dashboard to Monitor Data Quality Issues – Not Valid Range

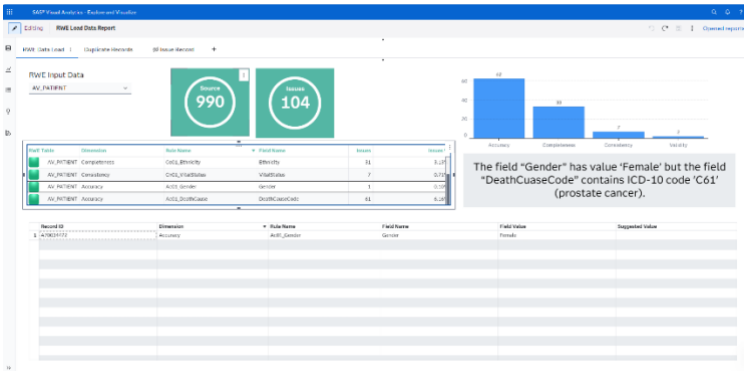

Figure 8 Dashboard to Monitor Data Quality Issues – Logical Error

The user can further expand the dashboard to support prioritization and impact assessment by highlighting which data quality issues have the greatest business impact. For example, issues in mandatory clinical outcome variables may be flagged as "critical," while minor formatting issues are marked "low severity." This helps teams prioritize remediation efforts in the time constraints. The user can also build heatmaps or bar charts to show the most common rule violations by table and by variables, which can help teams to identify if there is any trend and systematic error in data quality and can intervene accordingly. This approach can help organizations quickly detect systemic problems, such as recurring missing fields from a particular source system.

Decisions provide governance by tracking inputs, outputs, and execution logic. They also allow integration with other SAS Viya services or external applications, enabling continuous data quality monitoring as part of larger workflows, which we will discuss in the next section.

2. THE ANALYTICAL INVESTIGATION TIER

GO BEYOND BUSINESS RULE-BASED QUALITY CHECKS

Our example above is mainly rule-based validation, but there are many potential analytical layers that can enhance the rules-based approach.

- SAS Intelligent Decisioning can orchestrate not just rule-based validation, but also machine learning (ML) model outputs and even LLM-driven checks.
- Decisions can call predictive models registered in SAS Viya platform (e.g., anomaly detection). The model's output (e.g., probability scores or classification results) can then be used as additional quality indicators. For example, a model trained to detect outliers in patient lab results could flag unusual values for further review.
- Association analysis may highlight unexpected imbalances, or anomalous relationships worthy of further investigation.
- The data may be separated into "good" and "suspicious" records, and suspicious records can be compiled over time for investigation and pattern detection.

IMPACT OF MISSING DATA

A critical aspect of data quality assessment is evaluating the impact of identified issues on downstream machine learning models. Missing data, for example, can distort or bias models that predict patient outcomes, estimate treatment response or stratify patients into subgroups. Whether the values are missing at random or systematically is also important to understand.

An analytical layer can be built into the decision flow that simulates missingness and assesses its impact on model performance. Figure 9 shows simulated accuracy of a machine learning model under different missingness scenarios. It shows how the model steadily degrades as missingness increases[6].

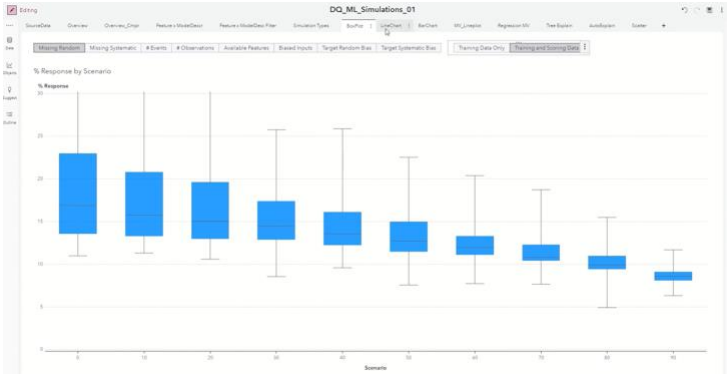

Figure 9 Visual Analytics Analysis of the impact of missing data on an unsupervised machine learning model

The view in Figure 10 compares the impact on accuracy of the model in multiple different simulations, from missing random data or missing systematic data, in training and scoring models. In this example, as is consistent with established findings, the impact of missing random data in a training data set is much lower than the impact of missing systematic data in both training and scoring data. This highlights how random gaps in training data may be tolerable, while systematic gaps, especially in scoring data, can significantly degrade model performance.

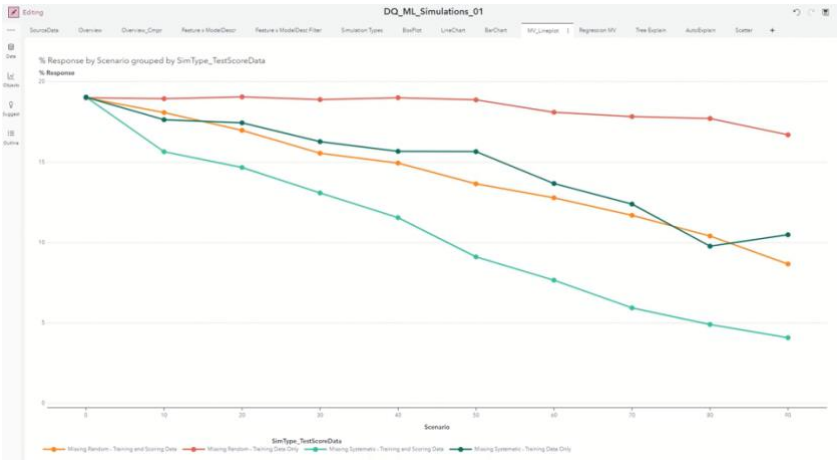

Figure x Visual Analytics Analysis of the impact of missing data on an unsupervised machine learning model

This quantification of impact on machine learning model performance metrics is a key step in setting acceptable thresholds for missingness – i.e. if 10% of values are missing and the model's AUC decreases by less than 0.02, the model still may be fit for use, but beyond 20% the degradation exceeds 0.02, breaching the predefined threshold. These thresholds can then be turned into go/no-go rules in the decision flow, support decisions on how to use the data and provide a level of confidence in the machine learning models[7].

When handling very large RWD data sets, this approach can be one of a number of "fitness-for-use" gates in an automated analytical pipeline, which then flags up that action needs to be taken.

3. THE ACTION TIER: SYNTHETIC DATA

SYNTHETIC DATA TO REMEDIATE DATA ISSUES

Once data quality issues have been surfaced and analyzed, the next step action. Options range from straightforward corrective actions to more advanced techniques, one of which is generating synthetic data.

Synthetic data is not yet a robust replacement for observational data or production data, although its use is being explored in synthetic control arms. It is designed to mimic the statistical properties and patterns of existing data and has immediate utility in improving the accuracy of machine learning models, via enhancing the quality of training and scoring data sets.

There are multiple methods for generating synthetic data. For example, generative adversarial networks (GAN) aim to produce as faithful a copy as possible of existing data, while synthetic minority oversampling technique (SMOTE) amplifies minority classes in the data for rebalancing.

Examples of it can effectively mitigate:

1. Imbalanced data. If certain groups are not well represented, synthetic data can be created to balance the dataset, using SMOTE.
2. A lack of rare events. Real-world datasets often fail to capture data on patients with ultra-rare diseases, for example, due to their rare nature. Synthetic examples of such patient data can enrich the dataset for the development of more reliable predictive models.
3. Insufficient data. If the limited or incomplete datasets do not provide enough information for training or testing models, synthetic data can be used to simulate real-world parameters and scenarios across the board to increase the data available.

SYNTHETIC DATA TO BROADEN DATA ACCESS

Poor data quality is often due to a lack of access and siloed data sets that lack interoperability. Synthetic data can aid in both scenarios. In healthcare and life sciences, HIPAA and GDPR regulations restrict access to sensitive personal information. With appropriate governance and guardrails, a data owner may permit the creation of synthetic data based on the original. This can open up access to analysis of the relationships in comprehensive and diverse datasets without exposing sensitive details. Synthetic data also has potential in federated data networks, where analysis is performed within each data node and aggregated without data ever leaving the source. Synthetic versions of the federated data

sources could enable more comprehensive data quality checks across the pooled data and therefore strengthen any federated analytics.

Regulatory bodies in Life Sciences have not yet established firm positions on the use of synthetic data, reflecting its nascent stage. The FDA has published draft guidance on the use of AI, including synthetic data, to support regulatory decision making[9]. The EMA has not yet released formal guidance and is running and evaluating pilot programs, including the use of synthetic data as a control arm in amyotrophic lateral sclerosis[10,11]. Risks that must be addressed include:

- Synthetic data can reflect inherent bias in the original data.
- There is a trade-off between accuracy and privacy, and it is challenging to create meaningful synthetic data sets that balance this effectively.
- Robust governance and documentation is essential at all stages.
- The quality of generated datasets must be high – and here our end-to-end data quality model can be used, with the generated synhtetic data passed through the flow after generation in a feedback loop.

INTEGRATING WITH OPEN SOURCE

SAS Viya offers both out-of-the box features in data quality and synthetic data and a flexible interface for highly sophisticated end-to-end analytic flows with a robust audit trail. The SAS language has its own strengths, but key is the ability to embed Python, R, and other open source code directly into workflows, enabling the best tools to be used for data preparation, modelling, and visualisation. This flexibility extends to sourcing data and algorithms from both internal work repositories and external web sources, allowing teams to enrich analyses with the latest innovations while maintaining enterprise-grade governance, security, and scalability.

Through SAS Viya integration with APIs, Decisions can call external LLMs as a node in the Workflow. LLMs used judiciously can support sophisticated quality checks on unstructured data. An LLM can assess whether a note or data attribute contains required information, or identify inconsistencies. Its output is then fed back into the Decision flow for automated quality scoring or routing. Our workflow, above, uses Python to call an LLM to assess the date format in the input data so that it can be efficiently corrected.

CONCLUSION

In summary, our tiered approach to RWD quality, which combines automated rule-based checks, advanced analytics, and synthetic data, offers a scalable, transparent solution for the evolving needs of life sciences. As the industry moves toward federated data networks and AI-driven evidence generation, robust data quality frameworks will be essential for trustworthy RWE. We invite collaboration across industry, academia, and regulators to refine these approaches and accelerate the adoption of high-quality RWD in clinical and regulatory decision-making, as the use of synthetic data evolves.

REFERENCES

1. Data quality framework for medicines regulation | European Medicines Agency (EMA)

2. Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products | FDA

3. Home page - simulacrum.healthdatainsight.org.uk

4. Intelligent Decisioning Software | SAS

5. SAS Studio: Working with Flows

6. Listening to your Data! Discover Relationships with Unsupervised Analysis Methods

7. Data Quality for Analytics Using SAS, Gerhard Svolba

8. Synthetic Data Will Save the Day: Fact or Fiction, Sherrine Eid, Sundaresh Sankarank, PAP_ET18.pdf

9. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products | FDA

10. Joint HMA/EMA Network Data Steering Group: Data and AI in medicines regulation NDSG workplan 2025-2028

11. Study control arms with AI support - Real4Reg

12. https://communities.sas.com/t5/SAS-Communities-Library/Orchestrating-amp-Governing-Large-Language-Models-in-SAS-Viya/ta-p/963185

CONTACT INFORMATION

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

Author Name: **Lisa Murch**
Company: SAS Institute
Address: 480 Argyle Street, Glasgow, G2 8NH, Scotland, United Kingdom
Email: lisa.murch@sas.com

Author Name: **William Yuanhao Kuan**
Company: SAS Institute
Address: Roppongi Hills Mori Tower 11th floor, 6-10-1 Roppongi, Minato-ku, Tokyo 106-6111, Japan
Email: william.kuan@sas.com