Building a Trustworthy RWD Analytics Platform: Key Considerations and Strategies

William Yuanhao Kuan, SAS Institute, Tokyo, Japan
Ashwini Reddy, SAS Institute, Mumbai, India

ABSTRACT

Real-world data (RWD) and real-world evidence (RWE) are gaining significant traction within the life sciences industry due to their potential to reduce research and development costs and accelerate patient access to innovative therapies. To facilitate the use of RWE in regulatory decision-making, the regulatory authority has issued several guidance documents outlining the appropriate use, benefits, risks, and methodological considerations for leveraging RWE in evaluating both the safety and effectiveness of medical products. Establishing an open and trusted research environment is essential for maximizing the value of RWD. Such an environment enables cross-functional collaboration and ensures that the generation of RWE is conducted with rigor, transparency, and confidence in its quality, reliability, and relevance.

We highlight the critical elements for building a robust RWE analytics platform and outline long-term strategies for utilizing RWD and generating regulatory-grade RWE supported by advanced analytical capabilities.

INTRODUCTION

RWD and RWE have become central topics in the healthcare and life sciences industry. According to the U.S. FDA, RWD refers to data related to patient health status and the delivery of healthcare that are routinely collected from various sources. RWE represents the clinical evidence about the use and potential benefits or risks of a medical product derived from the analysis of RWD. RWD encompasses information generated outside the setting of randomized controlled trials (RCTs), typically sourced from routine clinical practice, such as disease or product registries, insurance claims, and electronic medical records (EMR).

Why RWE such a prominent topic, and why does it matter to the healthcare and life sciences industry? In addition to RCTs, RWE offers the potential to reduce research and development costs and accelerate patient access to new therapies by harnessing the rich insights available from RWD. Increasingly, RWE is recognized as a critical source of evidence that can inform regulatory, reimbursement, and clinical decision-making. It can also guide the design and execution of clinical trials and support strategic program prioritization. This paper discusses the evolving RWE regulatory landscape in the United States, Europe, and Japan, as well as the key considerations and strategies for building a trustworthy RWD analytics platform for effective evidence generation.

USE CASES OF RWD/RWE

RWD refers to data generated in the context of routine clinical practice, such as electronic health records, insurance claims, and patient registries, unlike data obtained from randomized controlled trials. RWE is the clinical evidence derived from analyzing RWD. In Europe, the HMA/EMA Big Data Task Force uses the term “big data” to encompass diverse sources, including genomics, social media, and wearable device data. In Japan, the Pharmaceuticals and Medical Devices Agency (PMDA) RWD Working Group defines the use of RWD as the utilization of medical information databases. These databases include electronic medical records, diagnosis procedure combination (DPC) data from hospital information systems, claims for medical and dispensing fees, and disease registries.

RWE can play a valuable role throughout the entire drug development life cycle—from strategy development and clinical trial design to regulatory submission and post-marketing evaluation. RWD, for instance, can inform trial feasibility assessments, helping to refine study design and reduce the likelihood of costly protocol amendments. In addition, RWD can be leveraged to build external control arms in support of regulatory submissions. This approach is particularly useful in rare diseases, where enrolling enough participants for a RCT is often difficult, or when conducting such trials would be unethical, impractical, or impossible. In these cases, RWD-based external controls can help demonstrate the safety and effectiveness of a therapy relative to the standard of care. RWD can be leveraged in observational research to assess a product’s safety and effectiveness in actual clinical practice. A notable example is Pfizer’s use of RWE o support the approval of a label expansion for Palbociclib to include male breast cancer patients. The submission included findings from a retrospective outcomes study based on data extracted from EMR database. Despite some reservations regarding the robustness of the evidence, the FDA granted approval for the supplemental new drug application in April 2019.

In a review authored by Golnoosh Alipour-Haris et al.[1], the authors reviewed 85 relevant use cases between 2016 and 2022 and characterized the current use of RWE for regulatory submissions. The results show that in terms of therapeutic area, it consisted of 31 use cases in oncology and 54 use cases in non-oncology. The primary purpose of using RWE was for original marketing application approvals (59 cases), followed by label expansion (24 cases, and label modification (2 cases). Regarding the rationale for RWD use, the review found that 17 cases provided primary evidence, 42 cases were utilized to support single-arm trials, and 26 cases served as supplementary data for RCTs.

RWE REGULATORY LANDSCAPE

The utilization of RWD for post-marketing safety evaluations, such as safety surveillance studies, is relatively well established compared with its application in assessing treatment effectiveness. To promote a clearer and more consistent approach for regulatory decision-making, the U.S. FDA has released several guidance documents since 2018 outlining the appropriate use, potential benefits and limitations, and key considerations for generating RWE to support both safety and effectiveness assessments. The list of guidelines can be found in Table 1.

Table 1 List of RWE related guidelines from US FDA

Date Published	Title
July 2018	Use of Electronic Health Records in Clinical Investigations
December 2018	Framework for FDA's Real-World Evidence Program
September 2022	Submitting Documents Utilizing Real-World Data and Real-World Evidence to FDA for Drugs and Biologics
February 2023	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products
August 2023	Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products
December 2023	Data Standards for Drug and Biological Product Submissions Containing Real-World Data
December 2023	Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products
March 2024	Considerations Regarding Non-Interventional Studies for Drug and Biological Products
July 2024	Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products
December 2025	Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices

As of 2026/01/19

The guidelines encompass a wide range of topics, from fundamental principles to submission requirements, and provide criteria for evaluating RWD sources such as registries, EMR, and medical claims data. They apply to various health technologies, including drugs, biologics, and medical devices. The guidelines cover the activities such as expanding or updating an approved indication, generating comparative effectiveness or safety data, and serving as an external control arm.

In December 2024, Center for Drug Evaluation and Research (CDER) established a new Center for Real-World Evidence Innovation to coordinate, advance, and promote the use of RWD and RWE in regulatory decision-making across CDER offices. In December 2025, FDA states it will accept RWE without requiring that identifiable individual patient data for certain medical device marketing applications. This policy shift expands eligible RWD sources, enabling observational studies to leverage massive de-identified datasets.

In Europe, the Data Analysis and Real World Interrogation Network (DARWIN EU) serves as the cornerstone for more use of RWE. Through this network, regulators can access and analyze healthcare data from across EU member states, generating timely and reliable evidence on the use and impact of medicines.

In Japan, PMDA has placed greater emphasis on the application of RWD for post-marketing safety assessments. In 2021, PMDA released guidance on using patient registries in new drug applications. The PMDA increasingly views RWD as a critical tool for regulatory approval, particularly in the development of treatments for orphan diseases and pediatric populations where randomized controlled trials are often not feasible. The agency recognizes that utilizing RWD as an external control can provide vital supportive information, increasing the accuracy and scientific robustness of new drug applications that might otherwise rely solely on single-arm trials.

The FDA established its Framework for Real-World Evidence Program in 2018 to outline the rationale, scope, and approach for incorporating RWD into regulatory decision-making. Under the 21st Century Cures Act, the FDA's RWE

[1] Alipour-Haris G, Liu X, Acha V, et al. Real-world evidence to support regulatory submissions: a landscape review and assessment of use cases. Clin Transl Sci. 2024;17(8):e13903

Program is tasked with assessing how RWD can be used to evaluate product effectiveness—both to support approvals of new indications and to meet post-approval study obligations. The Framework highlights three primary considerations in regulatory use of RWD: (1) the fitness of the data for its intended use, (2) the suitability of the study design to generate scientifically valid evidence that addresses the regulatory question, and (3) the adequacy of study conduct in meeting FDA requirements. Overall, sponsors must demonstrate that the data and methods are robust and relevant to the regulatory objective, and that the entire evidence-generation process is transparent and reproducible.

Figure 1. Key Considerations for the Use of RWE in Submission

FEATURES OF A TRUSTED AND MODERN RWD ANALYTICS PLATFORM

This section outlines the essential characteristics of a reliable and modern RWD analytics platform, viewed through the lenses of Data, Methods, and Process. Such a platform is key not only to harnessing the full potential of RWD but also to supporting long-term strategies for its effective use.

DATA – RELIABILITY & RELEVANCE

When evaluating the suitability of RWD for a given regulatory purpose, both its relevance and reliability must be carefully assessed. To determine whether a selected data source adequately captures key elements such as exposures, outcomes, and covariates—and whether it includes a sufficiently large population—the platform should enable rapid data profiling through visualization tools and automated, advanced analytics. These capabilities allow users to explore data structure and quality, identify potential issues, and address them proactively before proceeding with analyses.

Equally important is the platform's ability to integrate and analyze multiple types of RWD originating from diverse sources and environments, both structured and unstructured data. By supporting interoperability and harmonization across heterogeneous data assets, the platform could empower users to leverage broader and more representative evidence. This diversity in data enhances the potential to uncover deeper insights into treatment patterns, outcomes, and patient heterogeneity, thereby strengthening the scientific and regulatory value of real-world evidence.

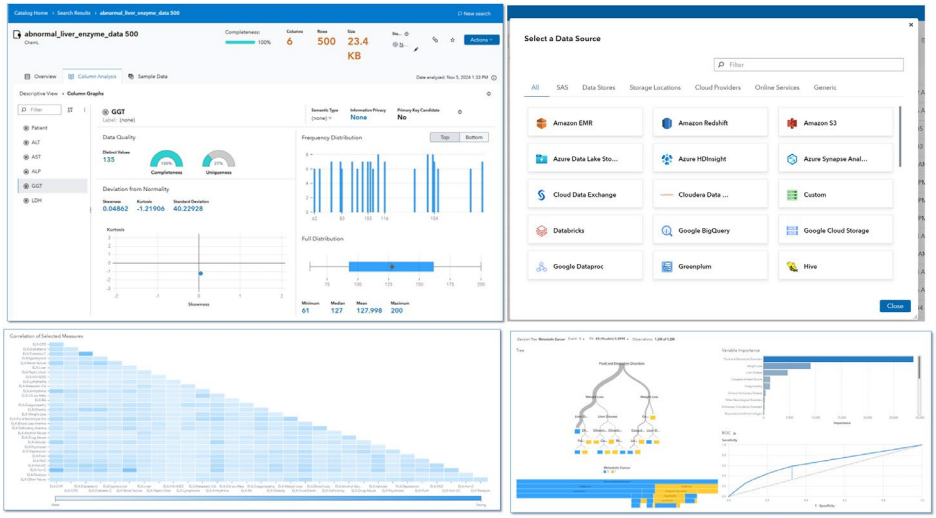

Figure 2. Automatic Data Profiling & Data Connectors

After identifying appropriate data sources, users must also confirm that the data is reliable. “Reliability” encompasses aspects such as accuracy, completeness, provenance, and traceability. Ideally, an analytics platform should employ a systematic data quality management process that clearly documents how data are transferred, transformed, and enriched, while also automating repetitive data preparation tasks. Such transparency enhances confidence in the curated and transformed data.

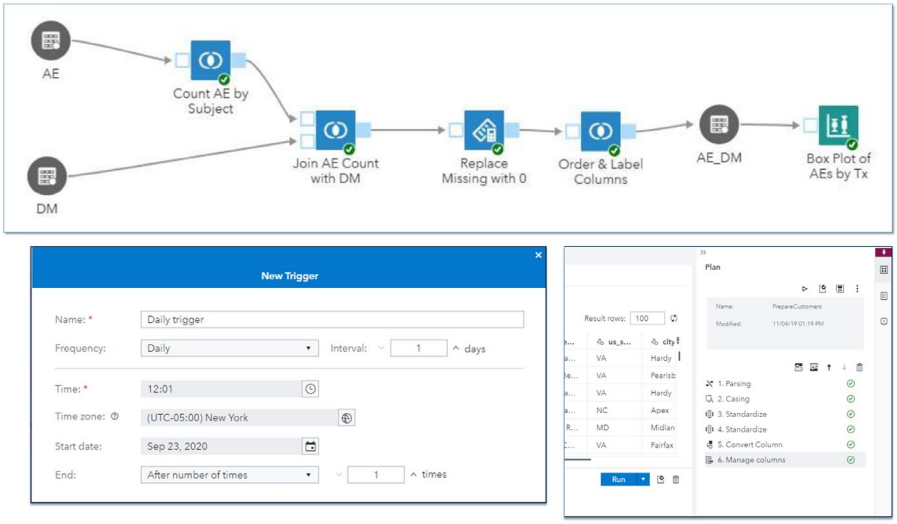

Figure 3. Transparent data flows and event trigger improve reproducibility

While electronic medical records, disease registries, and claims databases currently represent the primary sources of RWD, it is anticipated that more diverse data types, such as those generated from mobile devices, electronic patient-reported outcomes (ePRO), electronic clinical outcome assessments (eCOA), digital biomarkers, and medical imaging, will become increasingly utilized to generate deeper clinical insights. The FDA has indicated its intent to continue developing approaches to better leverage RWD from biosensors, ePRO instruments, and mobile or wearable technologies. Therefore, an ideal analytics platform should be capable of ingesting and processing large volumes of heterogeneous data and interfacing with various environments, including file formats such as PARQUET and JSON, covering different environments such as in-memory, databases (e.g., DuckDB), cloud (e.g., AWS, Azure, GCP), or streaming data sources.

METHODS – RELIABILITY & RELEVANCE

Besides using fit-for-purpose data, the analytics platform should enable the research team to demonstrate that the chosen study design and analytical methods are suitable and aligned with the regulatory questions being addressed. Moreover, the entire analytical workflow should be transparent and reproducible.

A RWE study should not be treated as a one-off project. By systematically managing the analytical assets generated across multiple studies, research teams can create standardized, reusable analytics templates, for example, patient phenotype algorithms, validated covariate definitions, or statistical analysis programs for safety and comparative assessments. This practice helps safeguard the organization’s analytic intellectual property. It also gives researchers confidence that they are applying programming code reviewed by quality control team and using appropriate methodologies and validated algorithms for RWD analysis. Over time, these reusable frameworks can support routine or simpler studies through standard templates, while more complex projects can employ customized programming. This is an approach consistent with practices adopted by DARWIN EU.

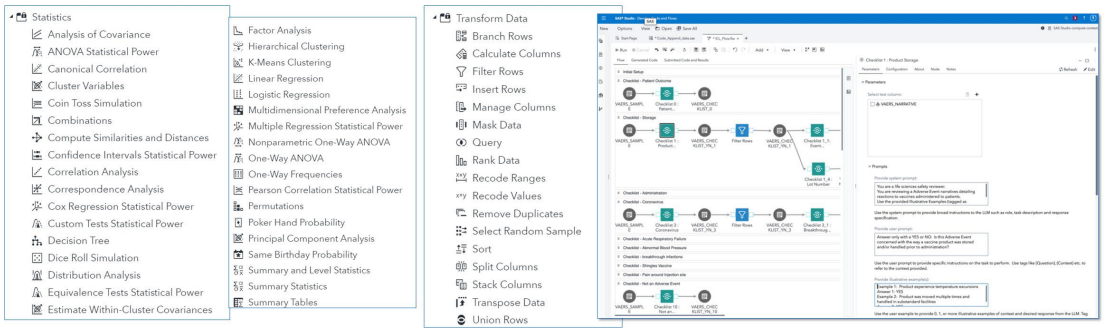

Figure 4. Reusable analytic templates/pipelines increase efficiency and reliability

PROCESS – RELIABILITY & RELEVANCE

Beyond the aspects of Data and Methods, ensuring transparency and reproducibility throughout the RWE generation process is equally critical, as these are fundamental for building stakeholder trust. The analytics platform should establish clear governance over data provenance and curation processes, and maintain complete lineage from raw data to final analyses, including results, execution logs, and audit trails. to support transparency, traceability, and reproducibility. Not only the table-level lineage, it is also important to have column-level lineage and impact analysis to show how a column was created/derived, where a specific column is used downstream, and how changes to that column will affect dependent tables, queries, jobs, and reports. Moreover, managing data, code, metadata, and analytic assets within a centralized environment further enhances traceability and oversight.

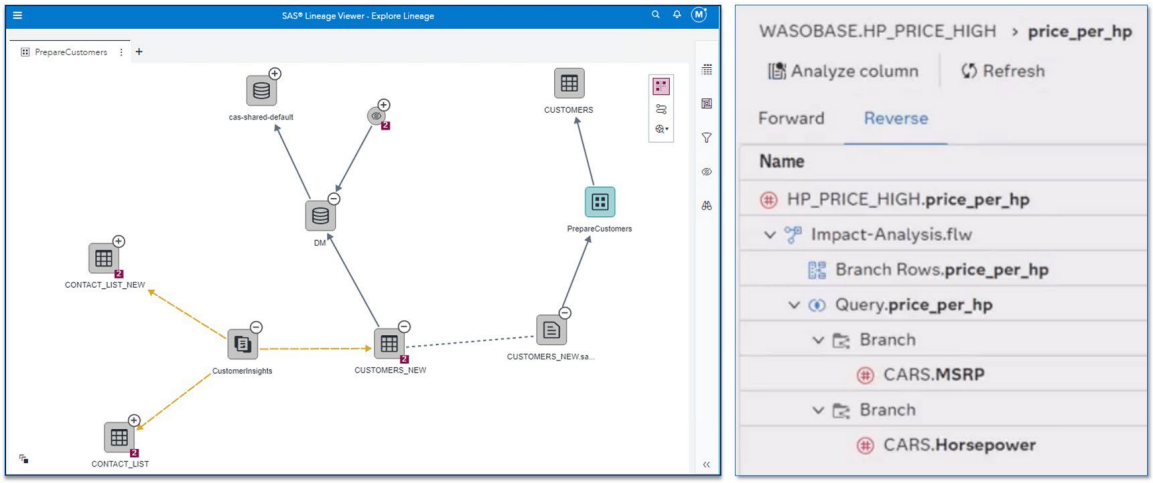

Figure 5. Data Lineage and Impact Analysis show traceability

With increasing attention on the use of AI in RWE, it is essential to have the capability to build end-to-end trustworthy AI practice. The use of AI and machine learning is rapidly expanding, supporting diverse applications such as treatment pattern identification, patient stratification, disease progression modeling, prediction of health outcomes, and extraction of insights from unstructured data sources like clinical notes or patient narratives. To ensure these AI-driven insights are credible, it is crucial to integrate explainable AI models, such as those employing LIME explanations and SHAP values, to help researchers interpret model outputs and confirm their clinical and methodological plausibility. In parallel, robust model governance frameworks should be implemented to document model development, validation, and deployment, while systematic fairness and bias assessments identify and mitigate algorithmic distortions that may compromise validity and equity. Continuous model performance monitoring is also essential to detect data drift over time, ensuring that AI models used for RWE use cases remain consistent, ethical, and scientifically rigorous across the full lifecycle of evidence generation.

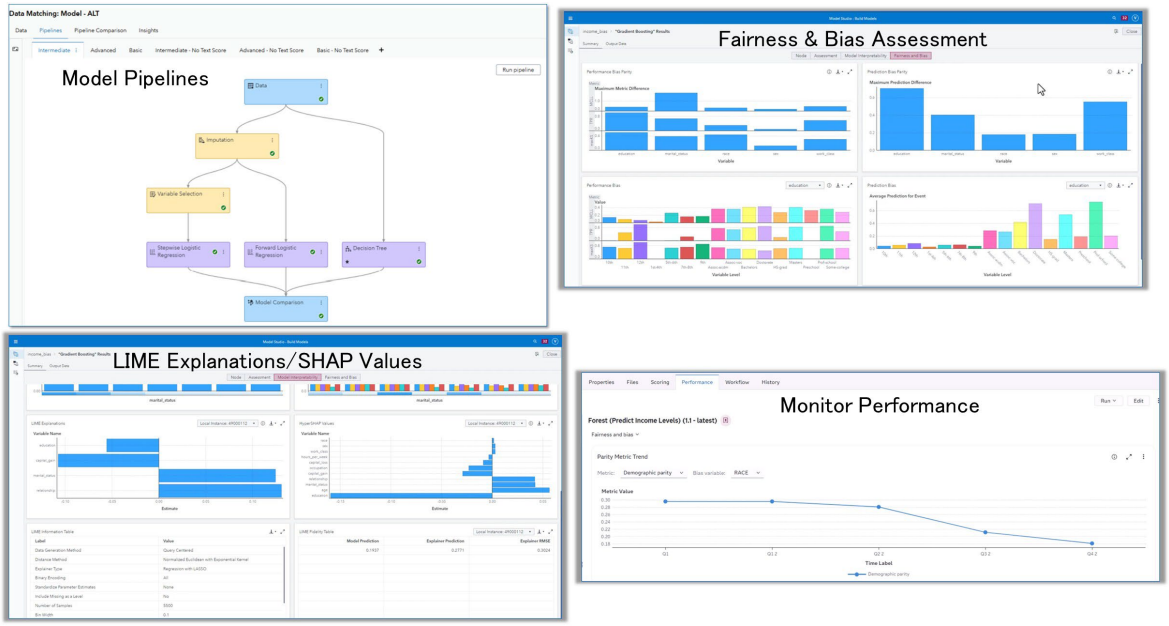

Figure 6. End-to-end trustworthy AI practices

FEATURES OF A TRUSTED AND MODERN RWD ANALYTICS PLATFORM – ADDITIONAL CONSIDERATIONS

This section outlines several additional factors that can help streamline the process of generating real-world evidence.

CLOUD NATIVE

A cloud-native platform capable of handling large-scale data is recommended. Such a platform provides the necessary computing power and storage, along with the scalability and flexibility needed to manage complex computational tasks. In-database or in-memory parallel processing technologies further enable high-performance analysis of large datasets. A cloud-native platform deployed in a secure, GxP-compliant hosting environment also reduces IT burden and allows geographically distributed research teams to access the platform under controlled conditions.

COLLABORATION

An analytics platform that supports seamless collaboration among users with diverse skill sets and from different organizational divisions can maximize the value of RWD assets. Business analysts can use interactive data visualization tools to explore data via no-code, while data scientists can perform more complex machine learning algorithms via low-code or yes-code. Shared access and low-code/no-code capabilities allow broader participation in data-driven decision-making and help break down organizational silos, driving greater efficiency and productivity.

CLASSICAL STATISTICS AND ADVANCED ANALYTICS

The ideal platform supports both traditional statistical methodologies, which are commonly used in clinical trial analyses and health economic outcome research for regulatory submission, and advanced analytics techniques such as AI/machine learning, computer vision, and natural language processing for exploratory analyses. As advanced analytics become more integral to regulatory submissions, the platform should ensure strong governance and transparency across model development, validation, and monitoring processes.

OPENNESS

An open platform supporting multiple programming languages (e.g., SAS, R, Python, Julia, Lua) allows users to work in their preferred languages and IDEs, fostering flexibility and improving talent recruitment and retention. Openness also means integration capabilities with third-party applications and the use of APIs to facilitate interoperability across systems, resulting in streamlining processes from data management to reporting.

Furthermore, openness should extend to integrating large language models (LLMs) and generative AI tools through protocols such as the Model Context Protocol (MCP). This enables seamless connections between analytics workflows and AI-assisted applications, allowing users to automate document generation, summarize analytic outputs, generate code templates, or build interactive dashboard by using natural language inputs.

UNIFIED, END-TO-END

A unified, end-to-end analytics platform that supports the entire analytics lifecycle is preferable. This eliminates the inefficiencies of switching between multiple tools. All essential tasks including data preparation, exploration, analysis, reporting, modeling, Generative AI and LLM integration can be performed within a single environment, promoting strong governance, consistent workflows, and effective collaboration.

CONCLUSION

In summary, while evidence generated from RCTs remains the gold standard for establishing the efficacy and safety profile of medical products, RWD/RWE can complement this process by offering broader context and more efficient approaches to address unmet medical needs throughout the drug development lifecycle.

Ensuring the reliability, relevance, and transparency of RWD sources, study design, and analytical methods is critical for regulatory submission use. Implementing a trusted and modern RWD analytics platform within a GxP-compliant environment enables organizations to effectively harness the value of RWD with greater confidence. To advance the regulatory use of RWD/RWE, the analytics infrastructure should offer robust computing capabilities and efficient data management to accommodate large, multi-source datasets. Moreover, the platform should support clear data provenance, systematic data quality control, and a governed, cross-functional collaborative environment that accelerates insight generation, and ultimately contributing to improved health outcomes and patient well-being.

CONTACT INFORMATION

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

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
Web: https://www.linkedin.com/in/williamyhkuan/

Author Name: **Ashwini Reddy**
Company: SAS Institute
Address: Maker Maxity, 3rd Floor, 4th North Avenue & 5th North Avenue, Bandra -Kurla Complex,
Bandra East, Mumba, India
Email: Ashwini.Reddy@sas.com
Web: https://www.linkedin.com/in/ashwini-reddy-93b98b89