Paper ET04

Accelerating Pharma Innovation: The Role of SAS Synthetic Data

Soundarya Palanisamy, SAS, Munich, Germany
Anand Phand, SAS, Pune, India

ABSTRACT

Drug development often depends on a limited pool of clinical data that is costly, restricted in use, and frequently insufficient or underrepresented. Leveraging synthetic data alongside real-world data can significantly accelerate drug development by reducing costs and enabling faster execution. Synthetic data uncovers hidden patterns within patient information while safeguarding privacy.
This session explores key considerations for generating synthetic data in a robust and governable manner, featuring SAS's new offering, Data Maker, as an example. We will review commonly used techniques such as Generative Adversarial Networks (GANs), SMOTE, and other emerging methods, along with approaches for assessing and validating data quality before application.

Additionally, we will discuss potential risks associated with synthetic data and strategies to enhance its acceptance in drug development. The session concludes with insights into health authorities' perspectives on the use of synthetic data in clinical trials, particularly within the context of drug development.

INTRODUCTION

Access to clinical trial data is often limited—data may be unavailable, difficult to access, or insufficient for robust analysis. Synthetic data, generated using algorithms rather than real-world collection, is designed to replicate the statistical properties and patterns of actual data.

By addressing privacy and data-sharing constraints, synthetic data enables secure and compliant collaboration. It augments small or incomplete datasets, improving trial design and strengthening analytical robustness. When real-world data is missing or of poor quality, synthetic data fills gaps and enhances overall data integrity. Ultimately, this accelerates drug discovery and research, helping deliver treatments to patients faster.

Industry surveys and our discussions with life sciences organizations reveal that the adoption of synthetic data in clinical trials is gaining momentum globally. Today, most use cases focus on improving trial design from the outset and simulating clinical trial data before actual data becomes available—streamlining processes and reducing time and effort.

Looking ahead, Gartner predicts that by 2026, up to 75% of companies across industries will leverage synthetic data, signaling strong momentum and confidence in its transformative potential (1).

ACCELERATING DRUG DISCOVERY AND CLINICAL TRIALS WITH SYNTHETIC DATA

Pharmaceutical companies use or want to use synthetic data because of two distinct challenges:

1. Either data is not available or difficult to access
2. Not enough data

DATA IS NOT AVAILABLE OR DIFFICULT TO ACCESS

Pharmaceutical companies often encounter challenges when essential clinical or patient data is unavailable or difficult to access due to privacy restrictions, regulatory hurdles, logistical issues, or the

absence of trial data. Synthetic data offers a practical solution by generating realistic datasets that enable researchers to design and test clinical trials, share information securely among stakeholders, and advance drug development—even when real-world data is not accessible.

Synthetic data safeguards privacy by replicating the statistical properties and patterns of real datasets without including any personally identifiable information (PII). This ensures organizations can use, share, and analyze data for research and testing without exposing sensitive patient details. Consequently, synthetic data supports compliance with privacy regulations and fosters secure collaboration across stakeholders.

NOT ENOUGH DATA

In many cases, especially rare diseases or specialized patient populations, there simply isn't enough data to support robust analysis or decision-making. Synthetic data helps fill these gaps by augmenting limited datasets.

SYNTHETIC DATA GENERATION USING SAS® DATA MAKER

SAS® Data Maker offers a low-code/no-code interface on the Viya platform, enabling users to quickly generate or augment synthetic data without writing code. Its intuitive UI provides built-in governance, auditability, transparency and trust features, making synthetic data generation accessible to both technical and non-technical users.

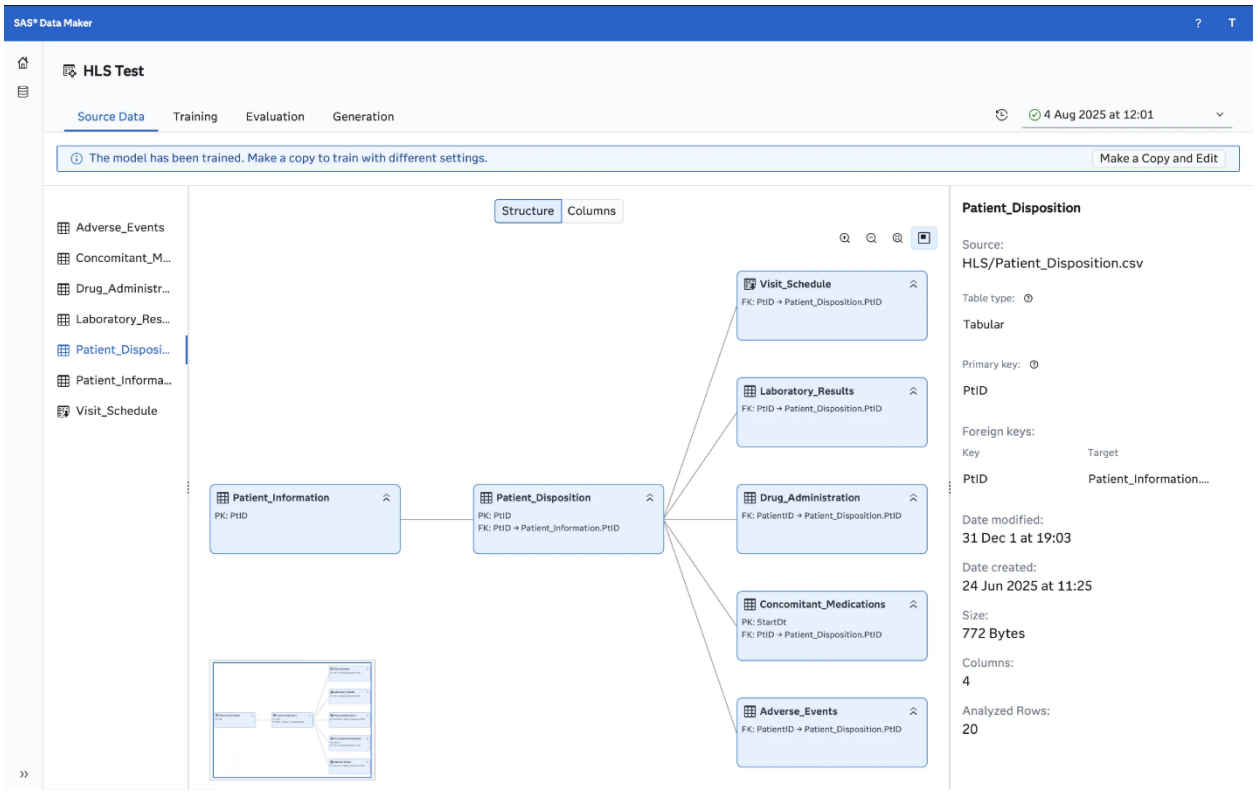

Figure 1: An example of generating synthetic clinical trials data using SAS® Data Maker

You can investigate synthetic data quality using various visual statistical evaluation metrics to verify that the synthetic data contains consistent patterns and insights as your real data.

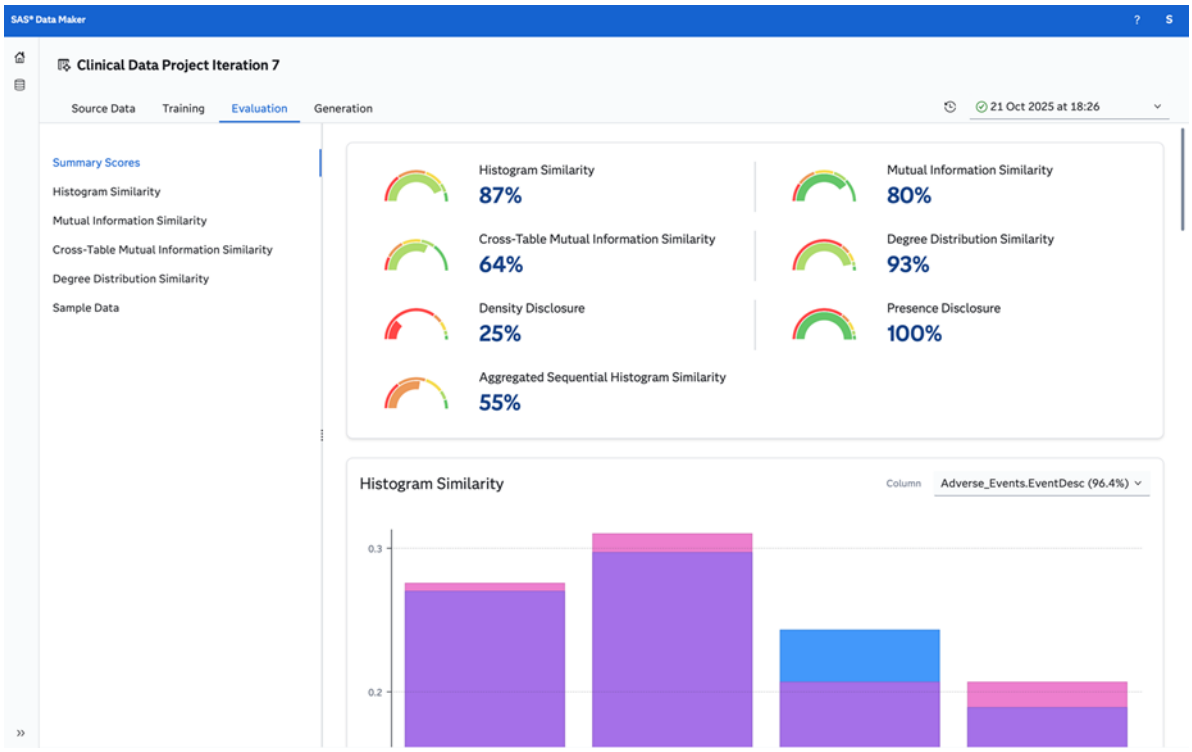

Figure 2: An example of visual statistical evaluation metrics in SAS® Data Maker

COMMON STEPS RECOMMENDED FOR GENERATION OF SYNTHETIC DATA

Figure 3 depicts a process for synthetic data generation. We consider the process to comprise of three phases – Plan, Prepare, and Produce.

In the **Plan phase**, the original dataset undergoes rigorous assessment, profiling and cleansing to identify inconsistencies, outliers, and sensitive, personally identifiable, and private variables, thereby establishing a reliable foundation for synthetic data generation.

The **Prepare phase** focuses on selecting appropriate synthetic data generation techniques. Two methods are in the forefront of real-world data applications: Generative Adversarial Networks (GANs) and Synthetic Minority Oversampling Technique (SMOTE). Each of these also has several variations. In this step also the generation parameters are specified, and the synthetic data generation model is trained.

Finally, the **Produce phase** involves generating and evaluating the synthetic data to validate the quality, realism and utility for intended use cases.

By following these steps, you can ensure that your synthetic data is of high quality, provides the right mix of similarity and privacy preservation, and is “fit for purpose,” i.e., serves its intended purpose effectively.

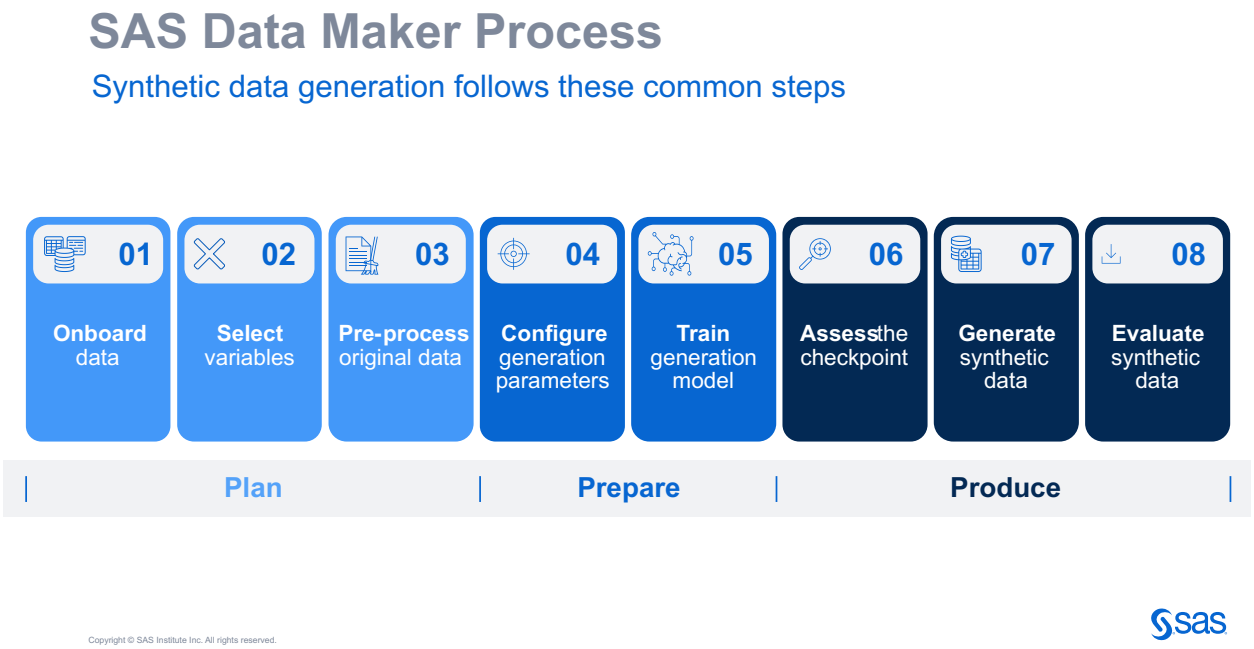

Figure 3: Synthetic Data Generation follows these common steps

METHODS AND TOOLS FOR GENERATING SYNTHETIC DATA

The idea of generating synthetic data is not new.
Synthetic data generation began in the early 1990s, with Rubin's 1993 proposal to use multiple imputation for privacy-preserving data and gained momentum around the turn of the century (3) . SMOTE (Synthetic Minority Over-sampling Technique) was introduced in 2002 to address class imbalance in datasets by generating synthetic minority samples (4). GANs (Generative Adversarial Networks) were proposed by Ian Goodfellow and his colleagues in their paper "Generative Adversarial Nets" in 2014 (5) and revolutionized the field by enabling highly realistic synthetic data generation, especially images and tabular data.

SYNTHETIC DATA GENERATION USING SAS® VIYA PROCEDURES

In SAS® Viya, synthetic data can be generated using advanced methods such as Generative Adversarial Networks (GANs) (6) and Synthetic Minority Oversampling Technique (SMOTE) (7) by leveraging built-in CAS actions and procedures like PROC TABULARGAN and PROC SMOTE. These procedures allow users to create realistic synthetic datasets by specifying inp tables, target variables, and output destinations directly in code, making it possible to automate and customize the data generation process for various research and development needs.

The SMOTE procedure enables you to use the synthetic minority oversampling technique (SMOTE) in SAS® Viya. SMOTE is an oversampling method that is used in machine learning to address the issue of imbalanced data sets by generating synthetic samples.

Example: PROC SMOTE in SAS® Viya

```
proc smote data=mylib.hmeq seed=123;
    input LOAN MORTDUE DEBTINC VALUE CLAGE CLNO YOJ/level=interval;
    input BAD JOB REASON DELINQ DEROG NINQ/level=nominal;
    output out=mylib.hmeq_out;
    sample numSamples=1000
          augmentvar=bad augmentlevel="1";
 run;
```

PROC TABULARGAN is a procedure in SAS® Viya that generates synthetic tabular data using generative

adversarial networks (GANs), enabling users to create realistic datasets that mimic the statistical properties of original data.

Example: PROC TABULARGAN in SAS® Viya

```
proc tabulargan        data=mylib.hmeq seed=123 numSamples=5 useGPU(device=0);
    input              value clage/level=interval;
    input              bad job/level=nominal;
    gmm                alpha=1 maxClusters=10 seed=42 VB(maxVbIter=30);
    aeoptimization     ADAM LearningRate=0.0001 numEpochs=3;
    ganoptimization    ADAM(beta1=0.55 beta2=0.95)  numEpochs=5;
    train              embeddingDim=64 miniBatchSize=300 useOrigLevelFreq;
    savestate          rstore=mylib.astore;
    output             out=mylib.out;
run;
```

HEALTH AUTHORITIES' PERSPECTIVE ON SYNTHETIC DATA IN CLINICAL TRIALS

Regulatory bodies such as the FDA and EMA are increasingly receptive to synthetic data in clinical trials, though their approach remains cautious and evolving.

FDA: Cautiously Optimistic

The FDA views synthetic data as a promising tool but stresses the importance of bias mitigation, rigorous validation, and proper evaluation methods. Key developments include:

- **Exploration & Guidance:** FDA Grand Rounds on Synthetic Data highlighted its potential and associated challenges.
- **Draft Guidance (Jan 2025):** Addresses AI-driven regulatory decision-making, including synthetic data generation.
- **Preclinical Focus:** Actively promotes non-animal testing alternatives under the FDA Modernization Act 2.0 (2022), authorizing advanced AI methods—such as GANs and language models—for IND applications.

In summary the FDA encourages synthetic data use, particularly in preclinical drug development, while emphasizing robust validation and bias control.

EMA: Cautiously Evaluating

Unlike the FDA, EMA has not issued formal guidance yet. However:

- **Strategic Workplan:** Synthetic data is under review in EMA's *Data and AI in Medicines Regulation to 2028* roadmap (Q4 2025).
- **Pilot Initiatives:** EMA participates in projects like **Real4Reg**, an EU effort leveraging AI and real-world data for regulatory decisions, including pilots on synthetic data for ALS treatments.

In summary EMA is in an exploratory phase, engaging in pilots and collaborations to assess synthetic data's regulatory potential.

CONSIDERATIONS FOR SYNTHETIC DATA

As described earlier, synthetic data offers numerous benefits when applied in drug discovery. However it is not a silver bullet. Some considerations include:

1. Bias and Fairness. Synthetic data can reflect **inherent** bias in original data and inadvertently introduce or amplify bias. This can affect the fairness of clinical trial outcomes and limit the generalizability of results

across diverse patient populations.
2. Privacy and Re-identification While synthetic data is designed to protect patient privacy, advanced re-identification techniques may still pose risks, and original data could be leaked.
Differential privacy is a powerful tool that introduces noise into data to protect individual identities (13).
3. Robust Data Governance is essential to ensure synthetic data is generated, managed, and used responsibly. This includes clear audit trails, transparency in data generation methods and parameters,
4. Regulatory Ambiguity Regulatory bodies such as the FDA are cautiously exploring synthetic data, but there is still ambiguity regarding its acceptance in clinical trials. The EMA has yet to issue concrete guidance, and sponsors must navigate evolving standards and expectations.
5. Data Quality and Utility Synthetic data must accurately reflect the complexity and variability of real-world patient data. Poorly generated synthetic datasets can compromise the reliability of trial results and undermine trust in findings.

Also domain expertise and human-in-the loop along with careful processes for synthetic data generation as mentioned earlier are essential for successful application of synthetic data.

CONCLUSION

Synthetic data holds significant promise for drug development, addressing challenges where real data is scarce, inaccessible, or insufficient. It has the potential to accelerate innovation and transform processes across the pharmaceutical lifecycle.

However, its use demands caution. To ensure trustworthy, auditable, and transparent outcomes, synthetic data should never be applied in isolation. A well-governed framework, supported by user-friendly, low-code tools that simplify generation without requiring programming expertise is essential.

Regulatory bodies remain cautious. While interest is growing, no major authority has issued definitive guidance on synthetic data for clinical trials. Adoption will require clear regulatory positions and robust governance.

In summary synthetic data could become a catalyst for pharma innovation, but its journey from promise to practice is still in its early stages.

REFERENCES

1. https://www.sas.com/content/dam/SAS/documents/infographics/2024/why-synthetic-data-is-essential-for-your-organizations-ai-driven-future-113845.pdf

2. (PDF) 30 Years of Synthetic Data

3. D. B. Rubin. 1993. Discussion: Statistical disclosure limitation. J. Off. Stat. 9:461–468

4. (Chawla et al., 2002) SMOTE: Synthetic Minority Over-sampling Technique

5. Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative Adversarial Networks. arXiv:1406.2661 [cs, stat].

6. PROC TABULARGAN: https://go.documentation.sas.com/doc/en/pgmsascdc/default/casml/casml_tabulargan_overview01.htm

7. PROC SMOTE: https://go.documentation.sas.com/doc/en/pgmsascdc/default/casactml/cas-smotesmotesample.htm

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

9. FDA Modernization Act 2.0: transitioning beyond animal models with human cells, organoids, and AI/ML-based approaches: https://pmc.ncbi.nlm.nih.gov/articles/PMC10617761/

10. FDA Grand Rounds on Synthetic Data: https://innolitics.com/articles/fda-grand-rounds-on-synthetic-data/

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

12. The EU project Real4Reg: unlocking real-world data with AI | Health Research Policy and Systems | Full Text

13. SAS Help Center: Differential Privacy

CONTACT INFORMATION

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

Soundarya Palanisamy
SAS Institute
soundarya.palanisamy@sas.com

Anand Phand
SAS Institute
anand.phand@sas.com