Paper TT01

Synthetic Data – the Game Changer in Drug Development

Soundarya Palanisamy, SAS, Munich, Germany
Anette Dalsgaard Jakobsen, SAS, Copenhagen, Denmark

ABSTRACT
Drug development relies on a limited volume of relevant clinical data which is restricted in use, costly to obtain and often insufficient and underrepresented. Augmenting real-world data with synthetic data accelerates drug development through lower costs and faster execution. Synthetic data surfaces valuable trends latent in patient data while avoiding compromise of privacy. This session provides an overview of considerations for robust and governable synthetic data generation, using a new SAS offering, Data Maker as an example. It provides an overview of available methods, such as Generative Adversarial Networks (GANs), SMOTE, and other emergent methods. We also cover quality assessment and evaluation of synthetic data prior to application. We also highlight potential risks from synthetic data and suggest how to increase acceptance in drug development. The session also touches upon the stance of health authorities towards synthetic data in clinical trials, particularly in the area of drug development.

INTRODUCTION

Clinical trials data is frequently either not available, difficult to access, or there is not enough data.

Synthetic data is artificially created using algorithms rather than being collected from real-world events. This type of data is designed to mimic the statistical properties and patterns of real data. Synthetic generated data can help solve privacy and data-sharing restrictions in clinical trials by enabling secure, compliant collaboration. It augments small or incomplete datasets, improving trial design and robustness. Synthetic data fills gaps when real-world data is missing or dirty, enhancing data quality.
It ultimately accelerates drug discovery and research, helping get treatments to patients faster.

Recent industry surveys and our own conversations with Life Science companies indicate that the adoption of synthetic data in clinical trials is accelerating among pharmaceutical companies worldwide. Currently we mostly hear in customer conversations synthetic data being used to help design better trials from the beginning and reduce effort and time. And for simulating clinical trials data before real clinical trial data is available to make the process more efficient and fast, once the data is available.

Looking ahead, Gartner predicts in 2026, up to 75% of companies across industries will be using synthetic data reflecting strong momentum and industry-wide optimism for 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 face situations where the necessary clinical or patient data is unavailable or difficult to access due to privacy restrictions, regulatory barriers, logistical challenges, or there is not yet data from a clinical trial. Synthetic data provides a solution by simulating realistic datasets, allowing researchers to design and test clinical trials, share data securely across stakeholders for collaborative research and trials and advance drug development even when real-world data cannot be obtained.

Synthetic data protects privacy by mimicking the statistical properties and patterns of real-world datasets without containing any personally identifiable information (PII). This allows organizations to use, share, and analyze data for research and testing without exposing sensitive patient or individual details. As a result, synthetic data enables compliance with privacy regulations and supports 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.

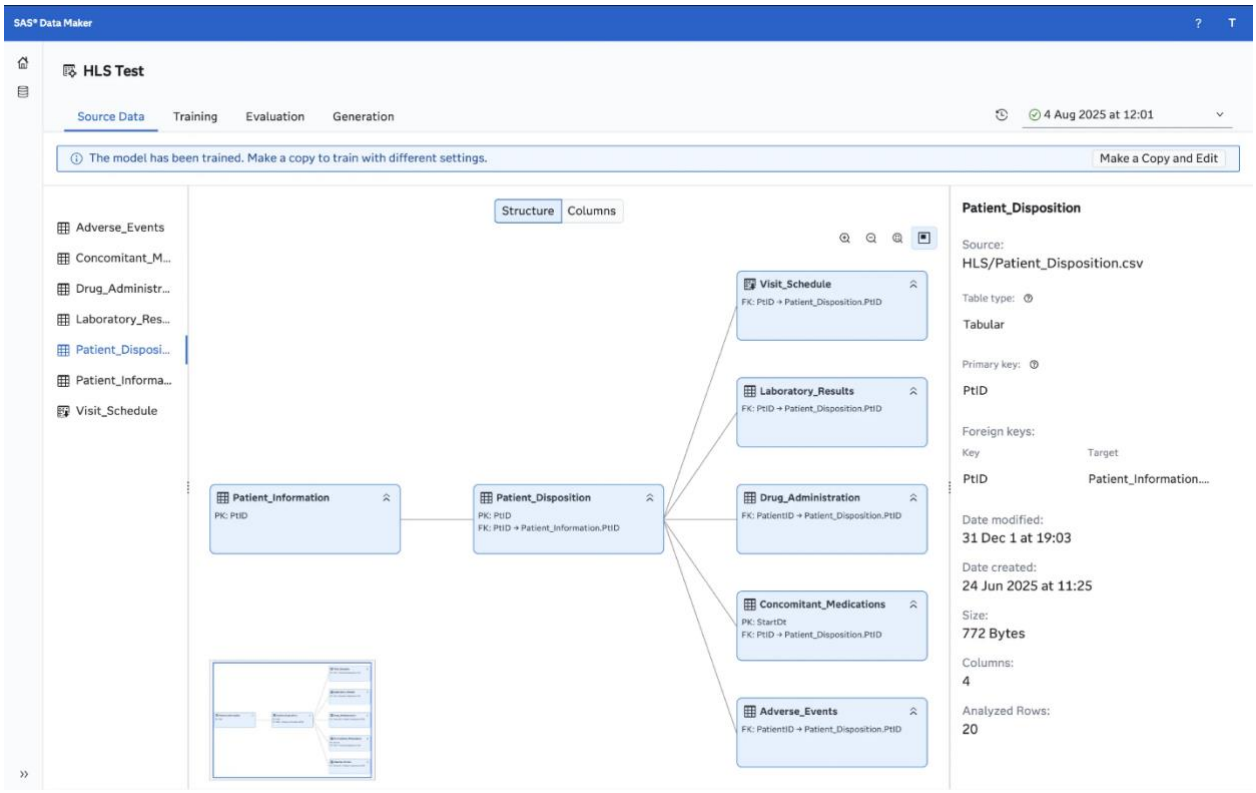

Figure 1: An example of generating synthetic clinical trials data using SAS® Data Maker

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.

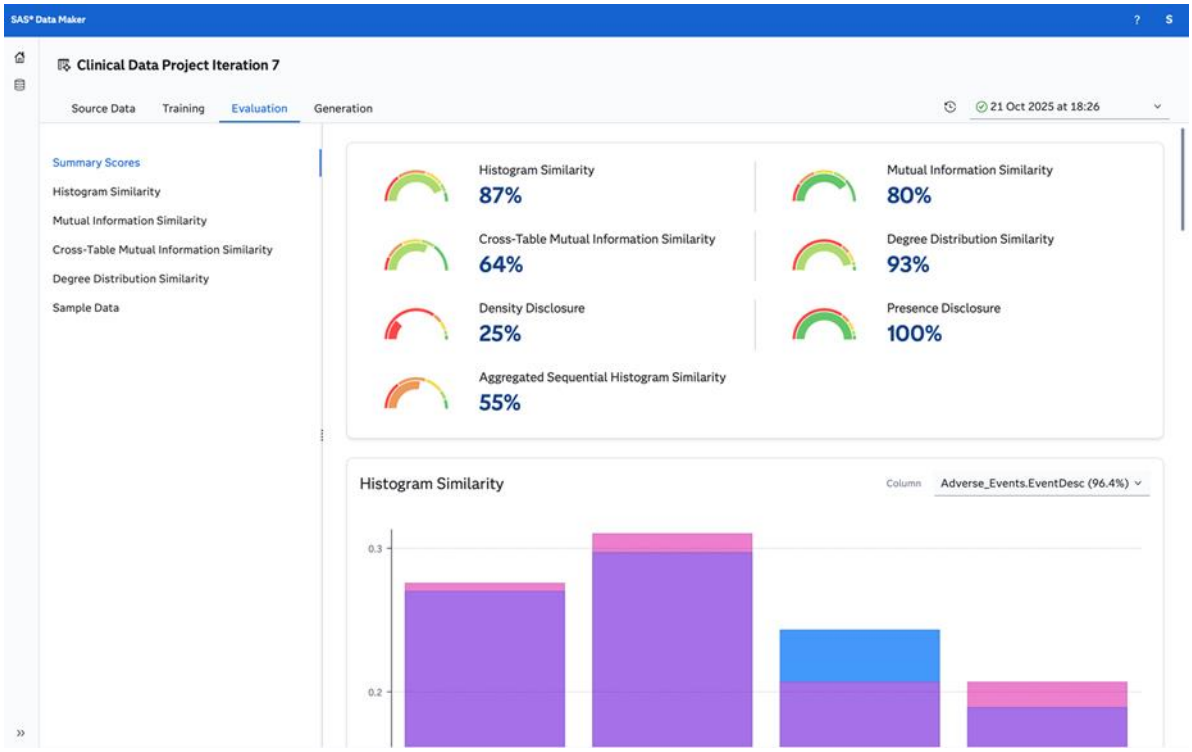

Figure 2: An example of visual statistical evaluation metrics in SAS® Data Maker

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.

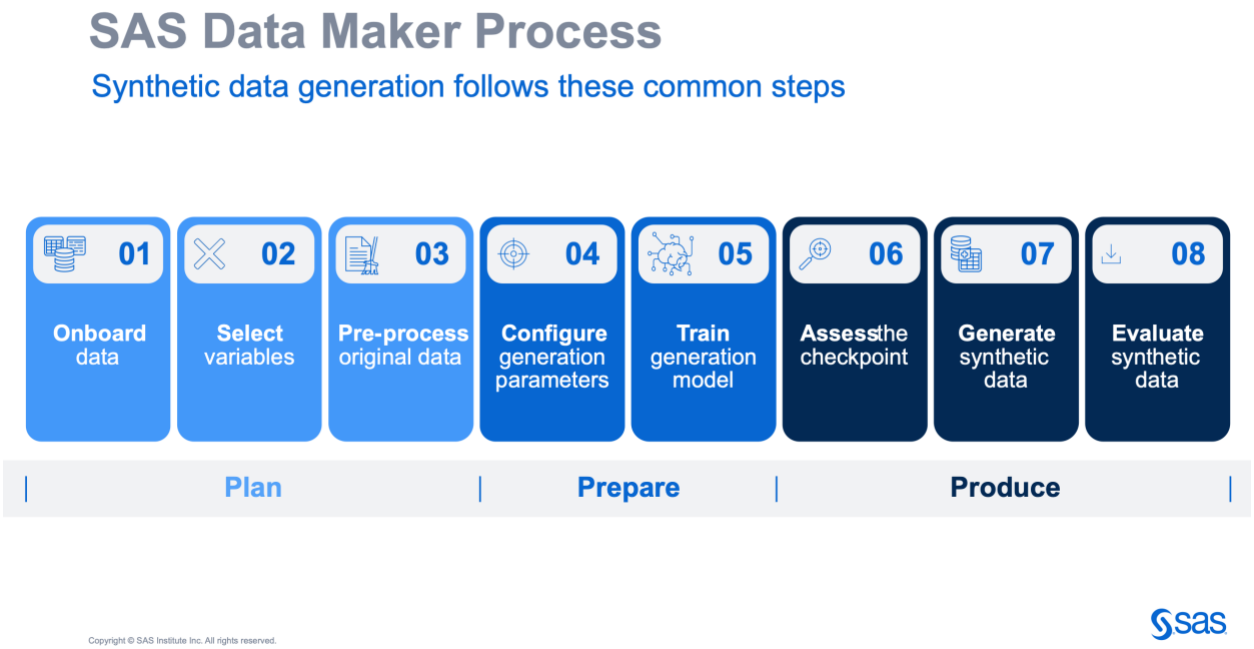

Figure 3: Synthetic Data Generation follows these common steps

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;
```

THE STANCE OF HEALTH AUTHORITIES TOWARDS SYNTHETIC DATA IN CLINICAL TRIALS

Health authorities like the **FDA** and **EMA** are increasingly open to the use of **synthetic data** in clinical trials, though their stance is cautious and evolving.

FDA : Cautiously Optimistic

The FDA is starting to explore the possibilities with synthetic data and sees it as a promising tool but also highlights addressing biases and proper evaluation and validation methods in their ***FDA Grand Rounds on Synthetic Data.***(10)

The FDA has published *draft* guidance (January 2025) on the use of **Artificial Intelligence (AI)** to support regulatory decision-making, which includes synthetic data generation (8)

In pre-clinical drug development FDA is actively encouraging the adoption of these methods as it will reduce the need for animal trials and the **FDA Modernization Act 2.0** (2022) (9) explicitly authorized the use of non-animal methods This comprehensive legislation permits the utilization of specific alternatives to animal testing, including cell-based assays, and advanced artificial intelligence (AI) methods, such as generative adversarial networks (GANs) and language models to support investigational new drug (IND) applications.

To sum up, the FDA is cautiously optimistic about synthetic data in drug development, encouraging its use—especially in preclinical settings to reduce animal trials—while emphasizing the need for rigorous validation, bias mitigation, and proper evaluation methods.

EMA: Cautiously Evaluating

EMA on the contrary has not yet released any formal guidance on synthetic data in drug development. Synthetic data is however one of the novel technologies being evaluated Q4 2025 in their **Data and AI in Medicines Regulation to 2028 workplan (11)**. EMA is engaging in pilot programs and international collaborations to evaluate synthetic data use cases. An example of a pilot project undertaken by EMA is the Real4Reg project is an EU initiative using AI to harness real-world data for regulatory decisions, including pilots on synthetic data and ALS treatments (12).

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 offers great potential in Drug Development to address the challenges with data that is either not available, difficult to access or not enough. Synthetic data might indeed be a candidate for a *Game Changer in Drug Development.*

However synthetic data must be used with caution and consideration to ensure trustworthy, robust, auditable, transparent and governed results. Synthetic data should not be used in isolation. A a well-designed process which helps teams generate synthetic data in a governed manner, assisted by user-friendly low-code components which do not require programming knowledge should be considered.

Also while opening up to the use of synthetic data most regulatory bodies still treat synthetic data with caution, and no major regulator has issued definitive guidance on synthetic data.

So synthetic data has *potentia*l to be *a Game Changer in Drug Development,* but it is still very early days in terms of delivering value in clinical trials and regulators need to clarify their stance

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/ru/pgmsascdc/v_065/casml/casml_smote_overview.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:

Anette Dalsgaard Jakobsen
SAS Institute
anette.dalsgaardjakobsen@sas.com

Soundarya Palanisamy
SAS Institute
soundarya.palanisamy@sas.com