



connect·share·advance

# Synthetic Data – The Game Changer in Drug Development

/ Soundarya Palanisamy, Anand  
Phand / 20. February 2026

[phuse.global](https://phuse.global)



# Agenda

- Introduction of Synthetic Data
- Use Cases and Demo
- Process and Methods for generating synthetic data
- Health authorities' stance on synthetic data in clinical trials
- Key takeaways



# Introduction of Synthetic Data

## **Synthetic Data Definition**

Synthetic data is artificially created data using algorithms rather than being collected from real-world events

## **Early Development**

The concept began in the 1990s with multiple imputation techniques for privacy-preserving data release.

## **SMOTE Technique**

SMOTE improved machine learning by generating synthetic samples to balance underrepresented classes.

## **GANs Breakthrough**

GANs revolutionized synthetic data by enabling realistic data creation through dual neural networks.



# Serious data challenges impede organizations from adopting and succeeding with AI.

## Data Privacy & Regulatory Compliance

Highly regulated industries must adhere to strict data privacy laws, making sharing or using real-world data difficult.

## Incomplete & Imbalanced Data

Real-world data often has gaps, missing fields, or imbalanced representations, creating biased or underperforming AI models.

## Cost of Data Collection

Collecting, cleaning, and labeling high-quality real-world data is time-consuming and expensive, particularly for large organizations handling massive datasets.

## Difficulty in Data Sharing

Organizations often face hurdles sharing data across internal teams and partners due to privacy and regulatory constraints.



# Synthetic data helps organizations overcome their data challenges quickly and securely.

## Simplifying Privacy & Compliance

**Anonymized Data for Safe Use:**  
Mimics real-world patterns without exposing sensitive information, ensuring compliance while enabling AI.

**Cross-border Data Sharing:**  
Allows easier data sharing across teams, departments, or even countries without breaching data residency laws.

## Augmenting & Balancing Data

**Fill Data Gaps:**  
Generate realistic samples to fill missing data, enriching datasets.

**Balance Data:**  
Balance rare events (like fraud or disease outbreaks), allowing AI models to learn effectively from all scenarios.

## Reducing Cost of Data Collection

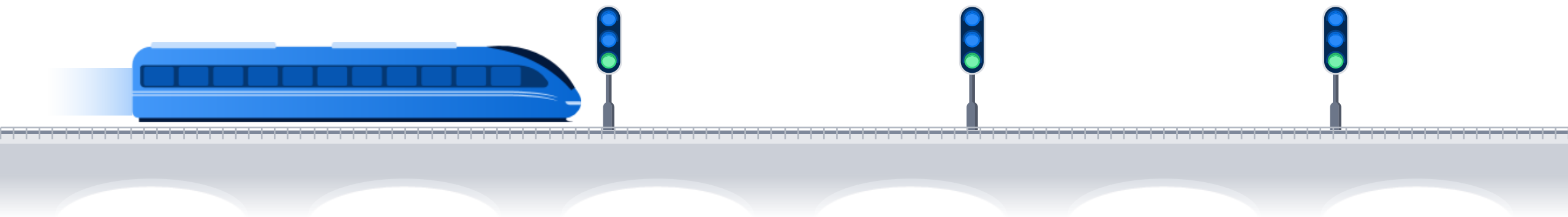
**Cost Efficient:**  
Can be generated programmatically, reducing the need for expensive data collection and manual labeling.

**Faster AI Deployment:** Accelerate data preparation, allowing your organization to move from concept to production more quickly.

## Enabling Data Sharing

**Facilitate Collaboration:** Share synthetic datasets that retain desired statistical properties without compromising privacy.

**Accelerate Innovation Ecosystems:** Ease of data sharing fosters collaboration across business units or with external partners, driving AI advancements.



# Accelerating Drug Discovery and Clinical Trials

Drug discovery and clinical trials often face challenges due to limited patient data availability and privacy restrictions. Synthetic data generation creates realistic patient datasets that preserve privacy while enabling faster, cost-effective research and development in drug discovery.

## The Problem



- Limited access to diverse patient data due to privacy regulations (e.g., HIPAA, GDPR).
- Small sample sizes and incomplete datasets slow down research and lead to biased results.
- Data-sharing restrictions hinder collaboration between pharmaceutical companies, research institutions, and healthcare providers.
- Data from clinical trials not available yet

## How Synthetic Data Helps



- Generates privacy-preserving synthetic patient datasets that mimic real-world data.
- Augments datasets to ensure diverse representation across demographics and conditions.
- Enables secure data sharing across stakeholders for collaborative research and trials.
- Simulating clinical trials data before real clinical trial data is available to make the process more efficient and fast, once the data is available.

## The Results



- Faster and more cost-efficient drug discovery processes.
- Improved clinical trial designs with better representation and robust data.
- Enhanced collaboration across institutions while maintaining compliance with data privacy laws.
- Design better trials from the outset – choosing the right trial setup to reduce wasted effort and time



# SAS® Data Maker

Your gateway to seamless synthetic data generation

[New project](#)

## Projects



### Retail

Time Series Forecasting

Date modified: Oct 27, 2025, 17:29

Modified by: Arpit Jain

Models: 2

Complete (2)

### Banking

Synthetic data generation for AML project

Date modified: Oct 27, 2025, 17:28

Modified by: Arpit Jain

Models: 1

Configuring

### HLS Demo Project

Synthetic Data Generation for Health & Life Sciences

Date modified: Oct 22, 2025, 14:33

Modified by: Arpit Jain

Training failed (6)

## New Project

Name your synthetic data project.

Project name:

Project description:

Create

Cancel

How to



### Overview

See an overview of SAS Data Maker.

[Watch video](#)



### Get Started

Learn about SAS Data Maker.

[View documentation](#)



### Provide Feedback

We would love to hear about your experience.

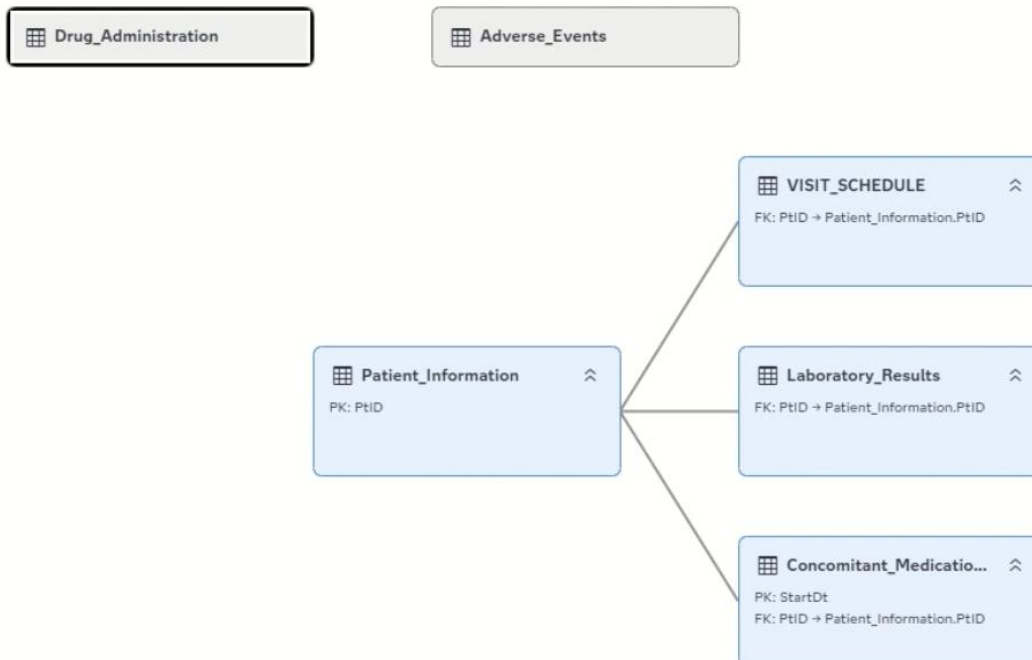
[Submit feedback](#)

AI Week 2025 Demo

Source Data   Training   Evaluation   Generation

- Adverse\_Events
- Concomitant\_M...
- Drug\_Administr...
- Laboratory\_Res...
- Patient\_Informa...
- VISIT\_SCHеду...

Structure   Columns



Drug\_Administration

Source:  
Demo  
Data/Healthcare\_and\_Life\_Sciences/Clinica  
L\_Study\_Data\_Sample/Drug\_Administration.  
csv

Table type:   
Tabular

Primary key:

Foreign keys:

| Key       | Target |
|-----------|--------|
| PatientID |        |

+ Add foreign key

Drug  
Date modified:  
Jan 1, 1, 00:53

Date created:  
Oct 27, 2025, 17:35  
AdminDate

Size:  
11.66 KB  
AdminTime

Columns:  
6

Analyzed Rows:  
246

# AI Week 2025 Demo

Source Data Training Evaluation Generation

🕒 Oct 27, 2025, 17:31

- Summary Scores
- Histogram Similarity
- Mutual Information Similarity
- Cross-Table Mutual Information
- Degree Distribution Similarity
- Sample Data

## Histogram Similarity

Column Adverse\_Events.EventDesc (76%)



AI Week 2025 Demo

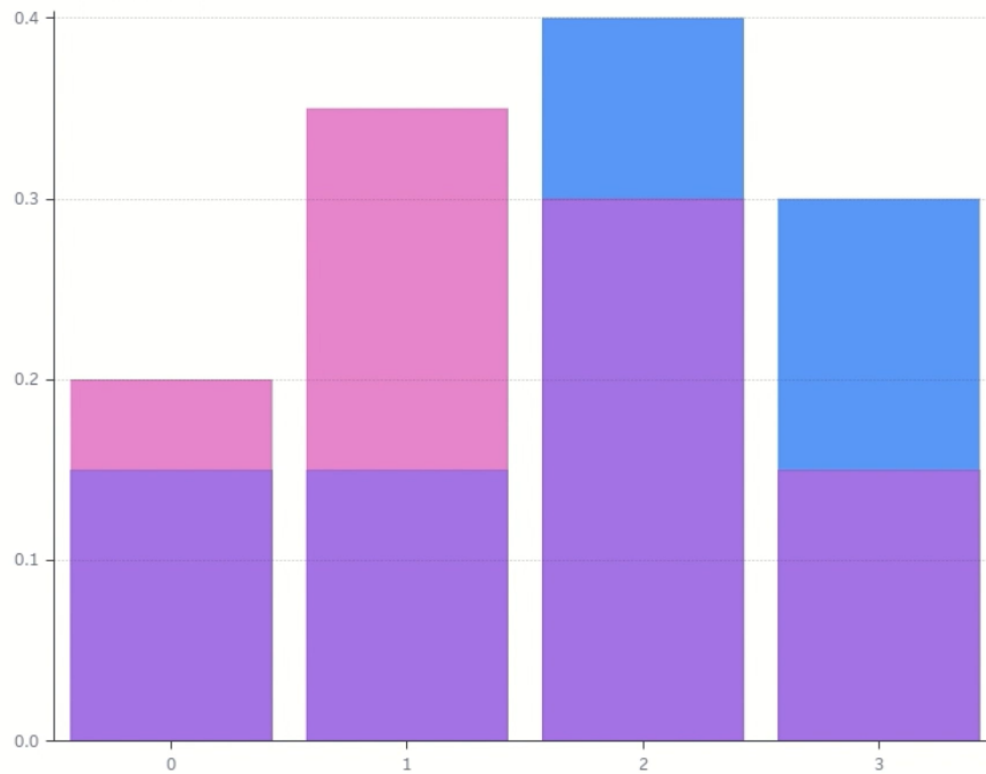
Source Data   Training   Evaluation   Generation

🕒 Oct 27, 2025, 17:31

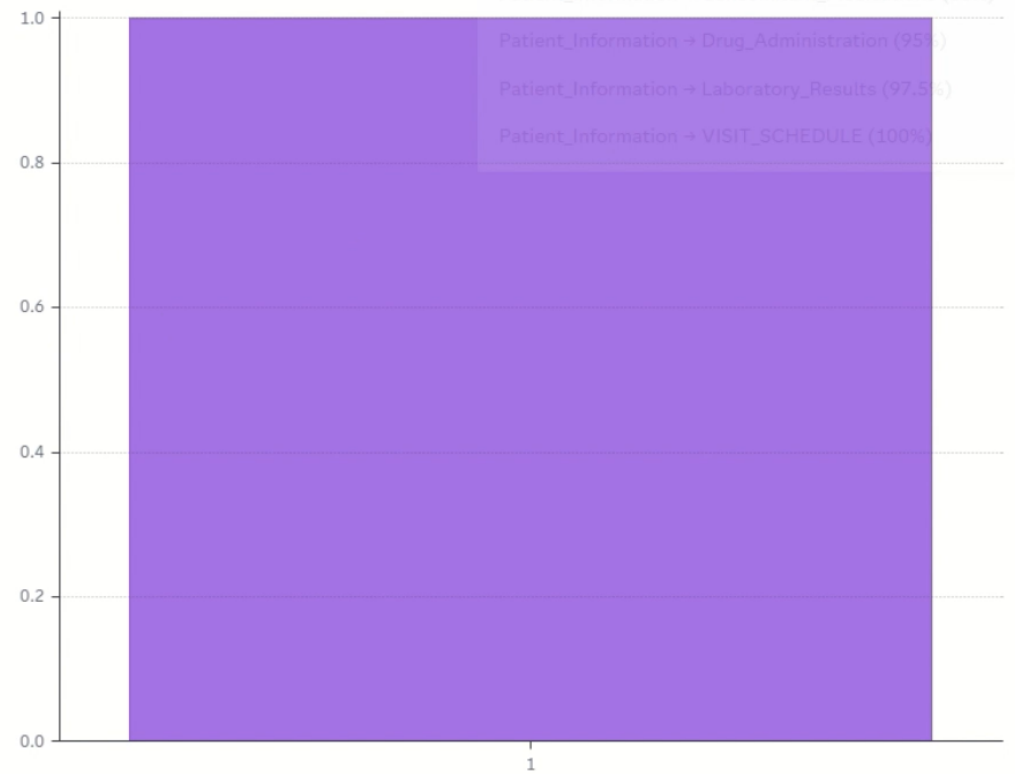
- Summary Scores
- Histogram Similarity
- Mutual Information Similarity
- Cross-Table Mutual Informatio...
- Degree Distribution Similarity
- Sample Data

Degree Distribution Similarity

Patient\_Information



Adverse\_Events



Tables   Patient\_Information → Adverse\_Events (88%)

- Patient\_Information → Adverse\_Events (88%)
- Patient\_Information → Concomitant\_Medications (85%)
- Patient\_Information → Drug\_Administration (95%)
- Patient\_Information → Laboratory\_Results (97.5%)
- Patient\_Information → VISIT\_SCHEDULE (100%)

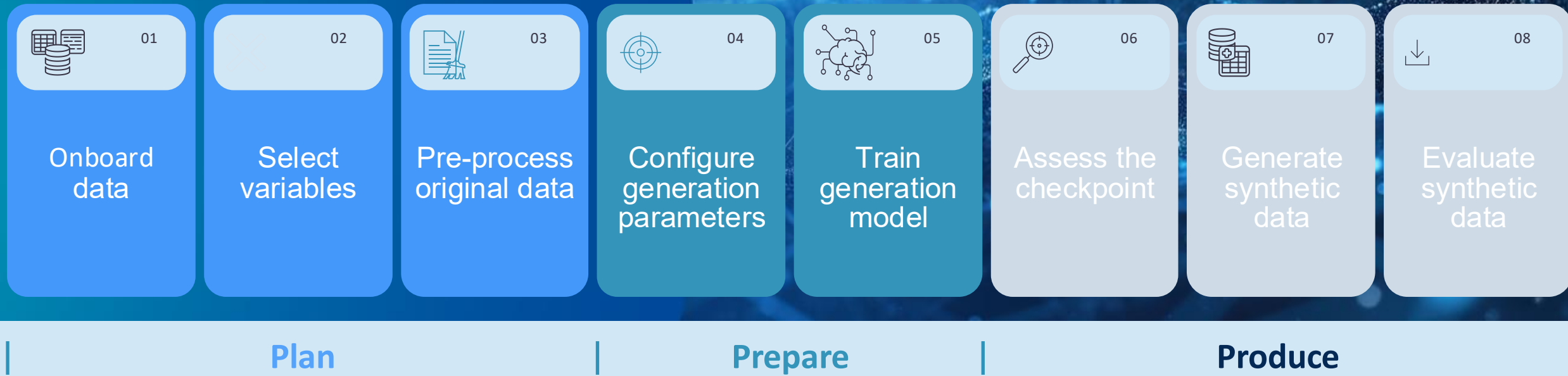


connect·share·advance

# Process



# Synthetic data generation follows these common steps





connect·share·advance

# Methods

# Common algorithms to generate synthetic data

connect·share·advance



General Adversarial  
Networks  
("GANS")

Deep-learning  
based generative  
models

Consists of 'generator'  
and 'discriminator'  
component

Can create many  
types of data

Computationally  
intensive

Useful  
for RWE

Can be run  
on SAS Viya

Nearest Neighbor  
Technique

Can reduce  
model bias

- Can improve the accuracy of  
ML models on imbalanced  
datasets

Can create  
unrealistic data  
points

Synthetic Minority  
Oversampling  
Technique  
(SMOTE)

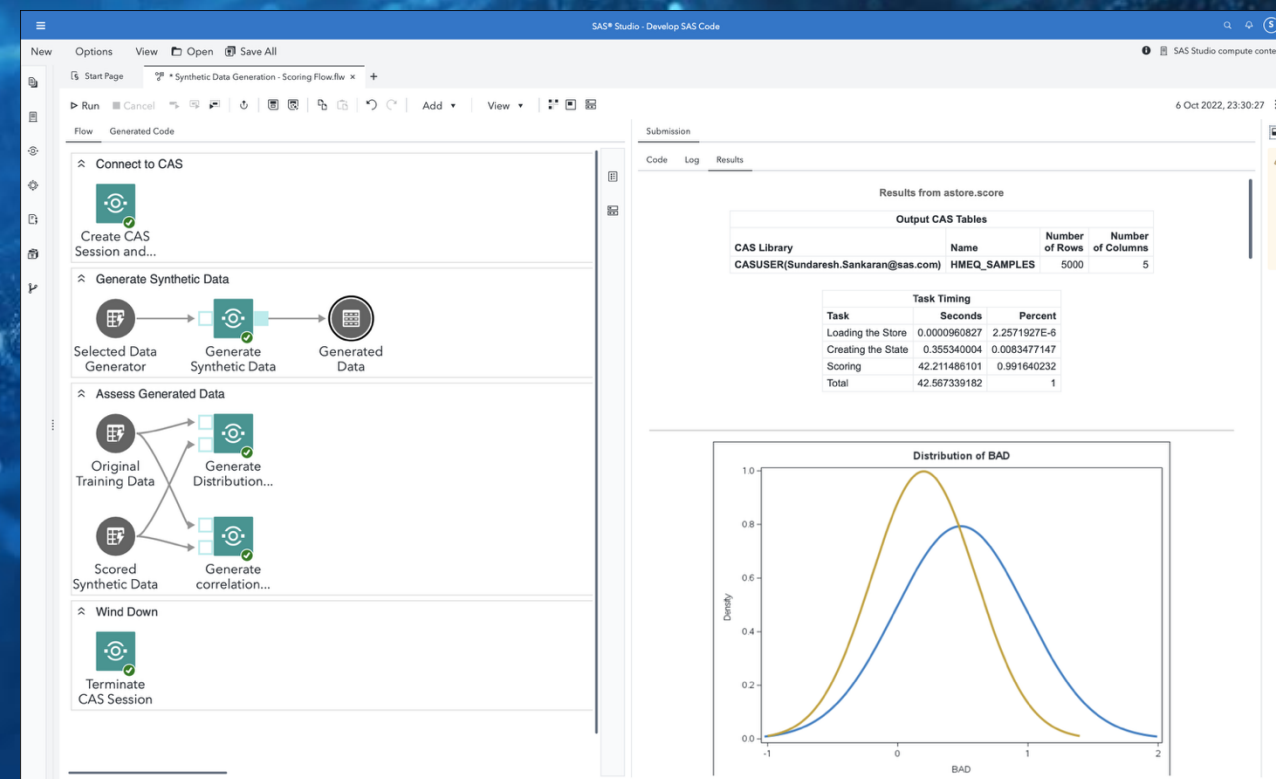
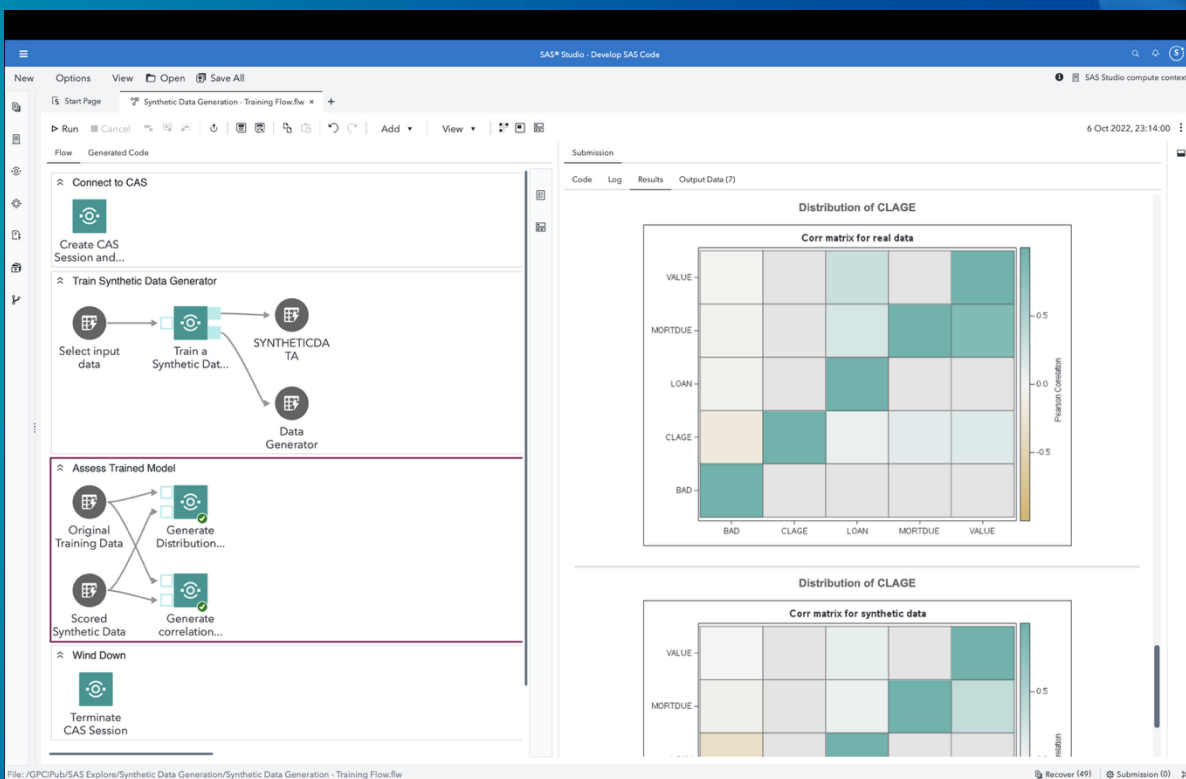


# Use SAS Studio Custom Steps on SAS Viya to train, score and assess synthetic data models (using tabularGAN)

connect · share · advance

Training a Synthetic Data Model

Generating Synthetic Data



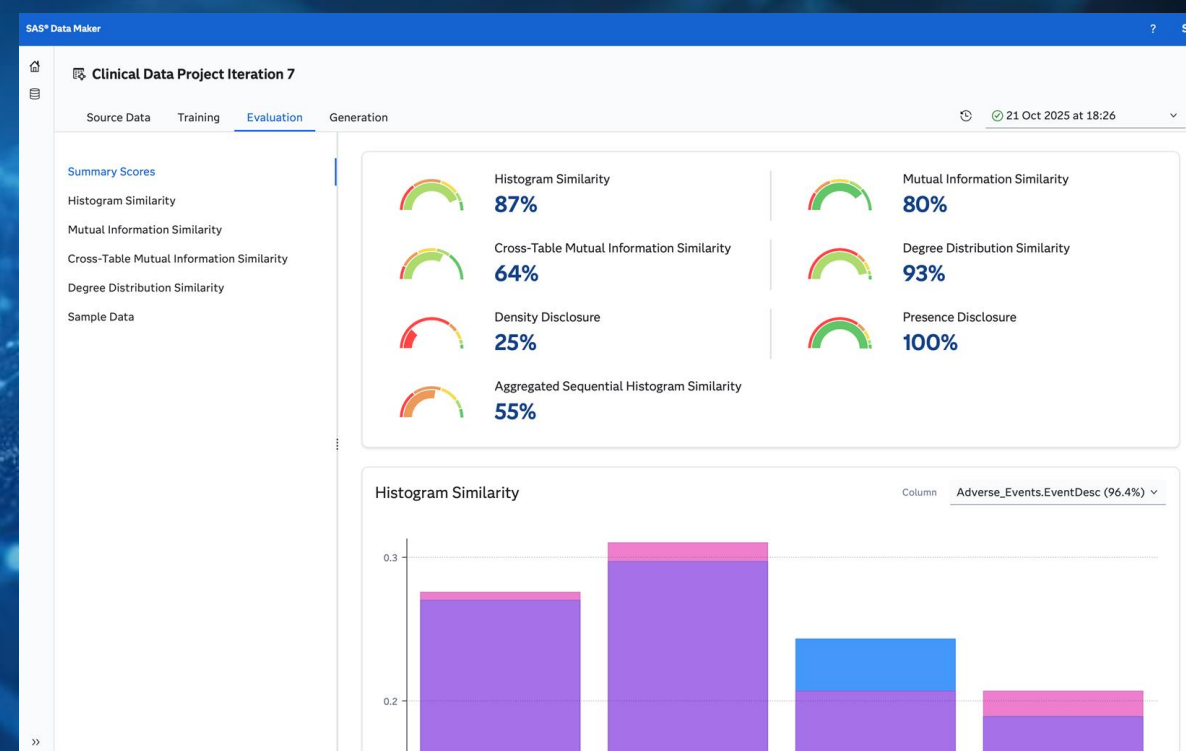
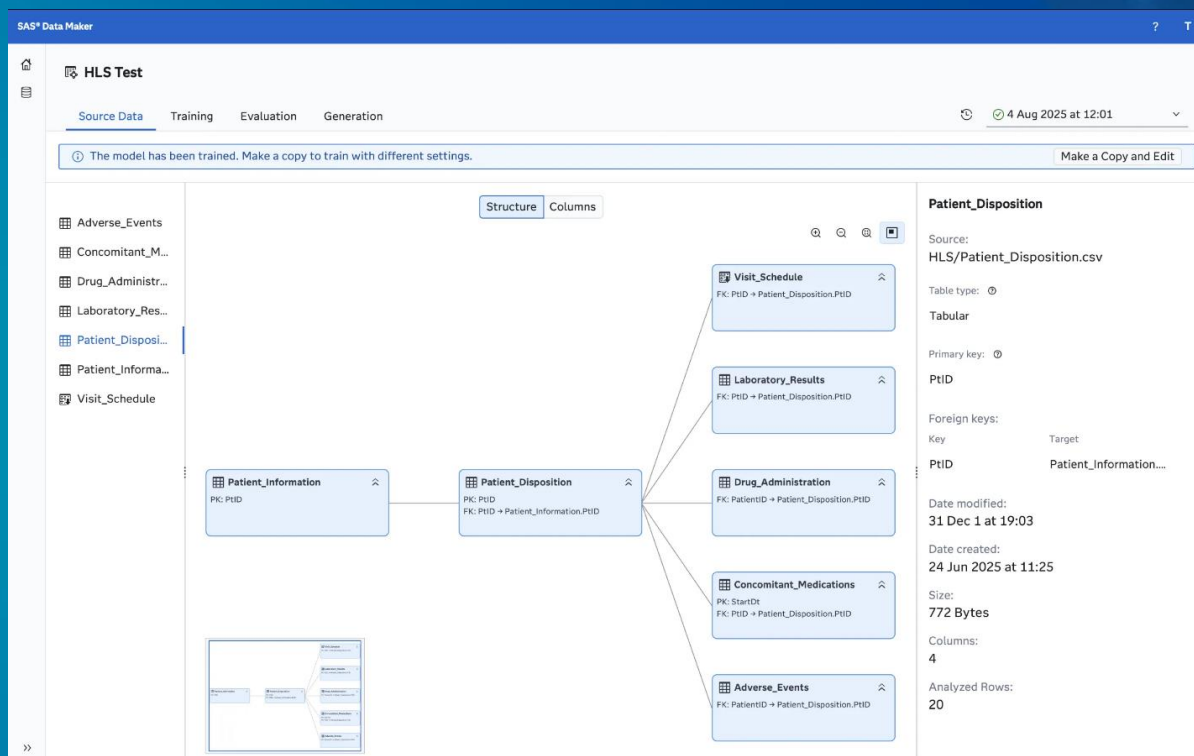
phuse.global

[sas-studio-custom-steps/Synthetic Data Generation at main](https://github.com/sas-studio-custom-steps/Synthetic-Data-Generation-at-main) · [sassoftware/sas-studio-custom-steps](https://github.com/sassoftware/sas-studio-custom-steps) · [GitHub](https://github.com)



connect·share·advance

# Synthetic Data Generation Using SAS® Data Maker



phuse.global

Screenshot For Illustrative Purposes



# Health authorities' stance on synthetic data

connect·share·advance

|          |      |                                                                 |
|----------|----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
|          | Exploring possibilities                                                                | Pilot programs to evaluate novel trial methodologies, including synthetic data                                                                     |
| Guidance | <i>draft</i> guidance (January 2025) on the use of <b>Artificial Intelligence (AI)</b> | not yet released any formal guidance on synthetic data in drug development.                                                                        |
| Focus    | Need for rigorous validation, bias mitigation, and proper evaluation methods.          | transparency, reproducibility and audit trails, ethical oversight, and risk validation, rigorous evaluation in synthetic datasets and applications |



## Key Takeaways

connect·share·advance

- Synthetic data has the potential to be a candidate for a ***Game Changer in Drug Development***
- Synthetic data should not be used in isolation, but rather in a **well-designed process**
- Health authorities are exploring **possibilities with synthetic data**



connect·share·advance

# Thank you!

[phuse.global](https://phuse.global)