Paper ET20

Challenges and Solutions for Generating Synthetic CDISC Data

Cat Briggs, SAS Institute Inc, Cary, NC
Sundaresh Sankaran, SAS Institute Inc, Cary, NC
Pritesh Desai, SAS Institute Inc, Cary, NC

ABSTRACT

Increased adoption of synthetic data, motivated by privacy protection, data sufficiency and simulation requirements, should align properly with CDISC data models, such as SDTM and ADaM, to benefit from standardization, improved quality and simplification. However, current synthetic data solutions are plagued by problems such as loss of referential and logical integrity, quality of results and inflexible methods. Using a reference dataset, we demonstrate challenges and pitfalls in generating synthetic CDISC data, and present likely solutions for consideration of multiple stakeholders across life sciences. This way, we expect to provide adequate preparatory guidance for implementing a new synthetic data practice and help successful planning and execution, overcoming challenges along the way.

INTRODUCTION

Synthetic data generation is the process of creating artificial datasets that mirror statistical characteristics, relationships, and distribution of real-world data. These datasets are produced using algorithms—simulations, statistical models, machine learning, or AI technologies—that learn patterns from existing data and generate new, realistic records that preserve key features but do not disclose actual sensitive information.

The practice is increasingly important in fields such as healthcare, finance, and research, where privacy concerns, regulatory restrictions, and limited data availability can hinder innovation. By generating synthetic data, organizations can safely test systems, validate models, and perform complex analyses without risking breaches of privacy or sensitive information. Moreover, synthetic data enables the creation of diverse scenarios for simulation and enhances machine learning by providing large, representative datasets—even when real data is scarce or inaccessible.

CURRENT STATE

Data privacy and compliance constraints greatly limit the use of real clinical datasets in life sciences. Regulatory frameworks like Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) strictly restrict the sharing and use of patient-level data, especially in research, development, and vendor testing environments. These laws are designed to protect sensitive information, but they also create significant barriers to data access, slow down innovation, and increase the complexity and cost of both data acquisition and processing. If organizations were to proceed with real data, even after their best efforts at adhering to regulation, they still face substantial risk of privacy breaches and non-compliance penalties, making it challenging to test systems, develop AI models, and perform comprehensive analyses in a secure, lawful manner. As a result, privacy and compliance constraints hinder the scalability and versatility of research workflows, ultimately limiting scientific progress.

APPROACH

This paper explores the generation of synthetic clinical datasets using the CDISC Pilot study as a foundational reference. Motivated by growing needs for privacy protection, data sufficiency, and simulation capabilities, our approach ensures alignment with industry-standard models such as SDTM and ADaM. The CDISC Pilot project—made possible through the sponsorship and data contribution of Eli Lilly—offers representative, standardized data that highlights both the opportunities and challenges in producing CDISC-compliant synthetic data. Our work specifically addresses pitfalls such as loss of referential and logical integrity, variability in result quality, and the limitations of current synthetic methods. By building on the CDISC pilot's legacy, we aim to deliver practical guidance and facilitate the successful planning and execution of synthetic data initiatives within life sciences, overcoming the most critical barriers to adoption.

OBJECTION HANDLING

Organizations typically raise objections to implementing a synthetic data solution (such as SAS Data Maker, which shall be used as an indicator application in this paper) around trust in synthetic data, regulatory acceptance, technical gaps, integration effort, and cost/ROI. These mirror broader concerns about generative AI and synthetic data in regulated industries.

- Trust and “fake data” perception
- Regulatory and risk objections

- Change management issues
- Cost, ROI and strategic fit
- Technical and feature-maturity concerns

The most effective way to overcome skepticism is to treat synthetic data like any other critical data asset: define quality, measure it rigorously against real data, and make the evidence visible to customers. Below is a practical playbook you can adapt to tools like synthetic data generators.

- Start with clear quality criteria and anchor conversation in concrete dimensions of quality, not in technology branding.
- Show quantitative validation against real data. Replace trust us with side-by-side evidence based on profile of real data using statistical methods like distribution, correlations etc.
- Prove downstream utility with model-based tests.
- Introduce governance bias and privacy checks.
- Communicate transparency and set scope with standard operating procedures.
- Run a targeted pilot with a joint sign-off. This shows that synthetic data quality is not taken on faith but demonstrated under their own governance and metrics.

SOLUTIONS

To achieve the above, the top priority of any synthetic data generation project is to include a data expert in the planning, preparing, Quality Control (QC), and use of the synthetic data. This way the data will be built for purpose with less irregularities or use issues. First, the data expert needs to create a reference map for the data. This includes Key IDs, relationships between tables and variables, and variables with repeated information. For example, Adverse Event tables have AEDECOD and AETER, which describe the event. These variables need to be connected in the algorithm to avoid nonreal code and term combinations in the synthetic data. Second, the data expert needs to check the synthetic data for other consistencies with the original dataset. For example, if code QC requires an expected distribution of sex, number of adverse events etc., the data expert must ensure the synthetic data has the desired distributions for its intended purpose.

Any synthetic data generation project must also have a data scientist run and manage the generative algorithms and models. This person needs to understand the requirements and nuances of creating synthetic data. Her role is to translate the data qualities set by the data expert to the algorithms for the best model. One instance is understanding and choosing the appropriate epsilon value for the differential privacy setting in a Private Bayesian (PrivBayes) model to balance accuracy and privacy of the synthetic data. The data scientist should also have access to a repository of models used for data generation. Within a robust process, the lineage of how the data is generated needs to be preserved and referenceable while that data is still in use.

Some common problems you'll encounter with their solutions are as follows:

1) The main purpose for creating synthetic CDISC data is to remove any sensitive and patient information but still retain the structure and patterns of the real data. A necessary step to avoid accidentally revealing true patient information is to use a technique called 'Differential Privacy'. Differential privacy minimizes the influence that an individual's data has on the trained generator. Smaller epsilon values lead to greater privacy guarantees, while larger epsilon values lead to greater accuracy. This is the trade-off the data scientist must make when considering the use of the synthetic data set. This technique can also only be used for Priv Bayes and similar algorithms, which may inform the chosen synthetic data generation process within teams. Additionally, the data scientist should use a density disclosure metric. This estimates the risk of an adversary somehow constructing a mapping from the synthetic data points to the real data points. This is a general privacy metric that should be used whenever privacy is a concern. It is recommended to use this metric whenever the data is to be shared with external third-parties or any untrusted group or individuals.
2) Synthetic data generation algorithms are probabilistic and not deterministic. This means the data scientist must explicitly label any connected or repeated information to ensure internally consistent data. This can be done by using a reference table that contains all the necessary connections, hierarchies, and repeated variables. Another way is using built in options within the software to label each connected variable in the data prep stage. For this case, we want the rows that have the "Placebo" ARM to also have "Pbo" ARMCD. In Tables 1 and 2, you can see that without an explicit input, the ARM and ARMCD variables could have conflicting information. The second table with only values on the diagonal (same designation for ARM and ARMCD) is necessary for correct synthetic data.

Table of ARM by ARMCD					
ARM	ARMCD				
	Pbo	Scrnfail	Xan_Hi	Xan_Lo	Total
Placebo	14	18	17	16	65
Screen Failure	17	22	16	22	77
Xanomeline High Dose	21	16	20	24	81
Xanomeline Low Dose	16	23	18	22	79
Total	68	79	71	84	302

Table 1: ARM and ARMCD for synthetically generated data without explicit internal logic input.

Table of ARM by ARMCD					
ARM	ARMCD				
	Pbo	Scrnfail	Xan_Hi	Xan_Lo	Total
Placebo	89	0	0	0	89
Screen Failure	0	72	0	0	72
Xanomeline High Dose	0	0	66	0	66
Xanomeline Low Dose	0	0	0	76	76
Total	89	72	66	76	303

Table 2: ARM and ARMCD for synthetically generated data with explicit internal logic input.

3) Without proper control, synthetic data may not match the distributions of the original variables. This can be the distribution of sex, race, body systems of adverse events, etc. The necessity of matching distributions depends on the intended use of the synthetic data. The data scientist has a few levers to pull to help match distributions. First, she can adjust the number of categories required for each variable to prevent some categories that could be dropped for a given variable. Secondly, include a proper table structure and reference to indicate sequential vs. tabular data. This forces the algorithm to follow the correct logic of the table type which often leads to more aligned data. For sequential tables, make sure the conditioning window size is large enough to account for dependent sequence of events as shown in Figure 1 and 2.

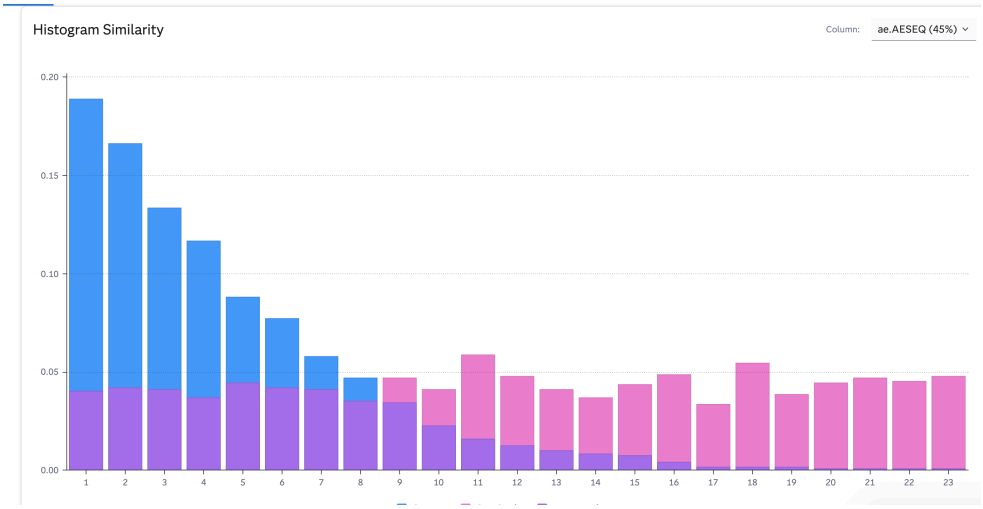

Figure 1: Histogram of distribution of Sequence Number in original and synthetic data without labeling it a sequential table.

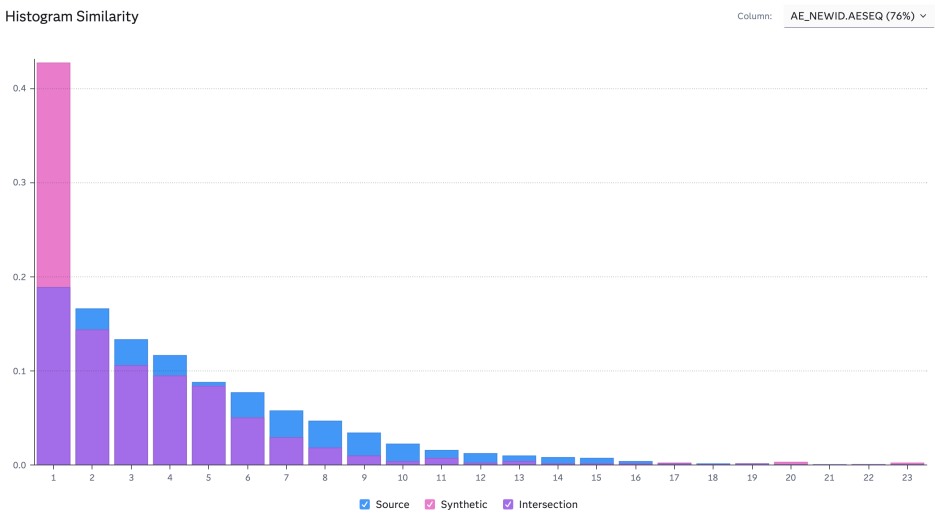

Figure 2: Histogram of distribution of Sequence Number in original and synthetic data with labeling it as a sequential table.

IMPLEMENTATION

To combine the expertise of the data expert and scientist, they need a reliable and trustworthy process to generate synthetic data. We recommend using established algorithms to quickly create synthetic datasets that accurately mirror real-world tabular and sequential data while protecting sensitive information and addressing known issues. Using tested software helps organizations innovate faster, reduce data acquisition costs, comply with privacy regulations, and fill gaps where non-sensitive data is scarce.

A good choice for software uses a combination of advanced statistical algorithms and machine learning models to generate synthetic data. The core methods should involve analyzing the original dataset to identify distributions, correlations, and statistical relationships, then replicating these patterns with algorithmically generated data. Key techniques may include:

- Tabular data simulation using learned distributions and correlations.
- Differential privacy algorithms to ensure no sensitive information is leaked.
- Support for synthetic oversampling algorithms, such as SMOTE (Synthetic Minority Oversampling Technique) for imbalanced data.
- Integration of newer generative AI models to construct complex, privacy-preserving synthetic datasets for domain-specific needs.

The tool should enable users to configure statistical methods, set magnitude/volume parameters, and evaluate data quality through metrics such as statistical overlap, correlation, and distribution comparison with the original data. This ensures the synthetic data is both realistic for modeling and safe for compliance use cases

The goal is to match the structure and variability of real clinical data, ensuring utility for analytics, AI model development, and regulatory testing without violating privacy or compliance rules. SAS Data Maker is an example of software with the features described above. It has a combination of advanced statistical algorithms and machine learning models to generate synthetic data. The technique involves analyzing the original dataset to identify distributions, correlations, and statistical relationships, then replicating these patterns with algorithmically generated data.

SUGGESTED ROADMAP

Knowing the common issues encountered above, how can the industry move forward to utilize synthetic data generation? Here, we suggest a few common steps for a smooth(er) synthetic data generation practice

- Understand the domains at which various levels of CDISC datasets operate (within their standards). For example, in SDTM, the DM domain can be carved out as a separate project to ensure high degree of data integrity within the domain, and then natural linkages to other domains are extended to connect them together. This ensures referential integrity while making synthetic data projects manageable.
- AI can learn patterns, but AI has problems learning "logic". Sometimes, there might be need for additional post-processing to ensure that logical, process driven relationships such as study visit dates aligning with visit numbers etc. are in order. Business rules and rule implementation frameworks aid this tremendously.
- Generate synthetic patient-level data for clinical trials, drug development, and research, replacing or supplementing real-world data for AI/ML model training and validation.
- Safeguard patient privacy by using synthetic data for test and development environments without exposing real patient records.
- Address data scarcity or regulatory limits by creating diverse and statistically accurate clinical case scenarios.
- Enable faster simulation, analysis, and reporting—especially for rare-disease studies, protocol testing, and compliance with data protection mandates.
- Evaluate and audit data quality with built-in, automated statistical metrics, ensuring that synthetic datasets maintain integrity for regulatory and research needs.

CONCLUSION

The adoption of synthetic data generation, when properly aligned with established CDISC models such as SDTM and ADaM, can significantly enhance privacy protection, data sufficiency, and simulation capabilities in life sciences. Our analysis, grounded in the CDISC Pilot study and facilitated by SAS Data Maker, demonstrates that while synthetic data presents significant promise, challenges related to referential integrity, logical consistency, and data quality must be carefully addressed. Leveraging standardized reference datasets and advanced generation tools enables organizations to overcome key limitations—streamlining processes, supporting compliance, and facilitating methodological innovation. Moving forward, successful integration of synthetic data practices will require a collaborative approach among stakeholders, with careful planning, validation, and adherence to industry standards to fully realize the benefits for research, development, and clinical analytics.

REFERENCES

1. Eli Lilly and Company. CDISC Pilot Study: Legacy Trial Data Contribution. Indianapolis, IN: Eli Lilly and Company. The availability of these valuable datasets and the advancements in clinical data standardization are directly attributable to Eli Lilly's sponsorship of the CDISC pilot study and their willingness to share legacy clinical trial data for this purpose. This foundational support accelerated the development of standards-based technology solutions in the health and life sciences industry.

2. CDISC. (2024). Study Data Tabulation Model (SDTM) Implementation Guide. Clinical Data Interchange Standards Consortium. Available at:
https://www.cdisc.org/standards/foundational/sdtm

3. CDISC. (2024). Analysis Data Model (ADaM) Implementation Guide. Clinical Data Interchange Standards Consortium. Available at:
https://www.cdisc.org/standards/foundational/adam

4. CDISC Pilot Project Team. (2023). CDISC Pilot Study: SDTM and ADaM Dataset Development and Implementation. Journal of Clinical Data Standards, 15(2), 45-62.

5. Hume, S., Aerts, J., Sarnikar, S., & Huser, V. (2016). Current Applications and Future Directions for the CDISC Operational Data Model Standard: A Methodological Review. Computers in Biology and Medicine, 89, 423-433.

6. CDISC. CDISC Pilot Study Data and Documentation. Available at: https://www.cdisc.org/kb/examples/adam-basic-data-structure-bds-using-cdisc-pilot-data

7. SAS Institute Inc., Eid, Sherrine and Sankaran, Sundaresh, "Synthetic Data: A Process Paradigm", Phuse US

connect 2025, https://www.lexjansen.com/phuse-us/2025/as/PAP_AS18.pdf

8. SAS Institute Inc. (2025). SAS Data Maker. Official product page. Available at: https://www.sas.com/en_us/software/data-maker.html

9. SAS Institute Inc. (2025). SAS Data Maker User's Guide. Documentation. Available at: https://documentation.sas.com/?cdcId=sdgcdc&cdcVersion=default&docsetId=sdgug&docsetTarget=titlepage.htm

10. SAS Institute Inc. (2025). SAS Data Maker Overview Demo. SAS Video Portal. Available at: https://video.sas.com/detail/video/6373908036112/sas-data-maker-overview-demo

11. SAS Institute Inc. (2025). Synthetic Data Output - SAS Help Center. Documentation. Available at: https://documentation.sas.com/doc/en/sdgcdc/v_001/sdgug/n19owjhzap06y4n1iepp442jpcwm.htm

12. Perplexity AI. (2025). Comet Assistant [AI research and writing tool]. https://www.perplexity.ai/

ACKNOWLEDGMENTS

This paper was developed with research and writing assistance provided by Perplexity AI (Comet Assistant), a browser-based AI system for literature review, information synthesis, and content generation.

CONTACT INFORMATION

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

Author Name: Cat Briggs
Company: SAS Institute, Inc.
Email: cat.briggs@sas.com

Author Name: Sundaresh Sankaran
Company: SAS Institute, Inc.
Email: Sundaresh.sankaran@sas.com

Author Name: Pritesh Desai
Company: SAS Institute, Inc.
Email: pritesh.desai@sas.com