Navigating Opportunities and Challenges in Synthetic Data Generation for Biomedicine: Insights from the SYNTHIA Project

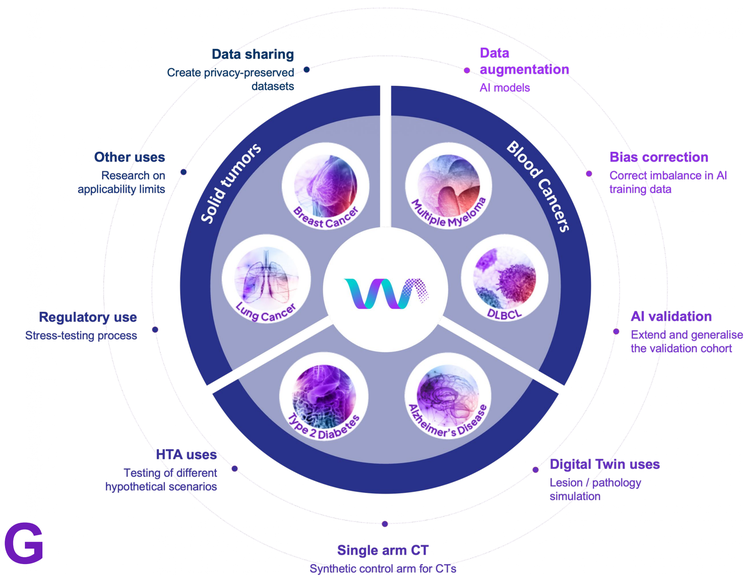
Patricia Alonso de Apellániz [1], Aleksandar Babic [2], Vibeke Binz Vallevik [2,3], Oscar Brück [4], Gastone Castellani [5], Chiara Chianese [6], Davide Cirillo [7], Vanda Czipczer [8], Saverio D'Amico [9,10], Mattia Delleani [9], Matteo Della Porta [10], Nicolas Derus [5], Chong Duan [11], Styliani-Christina Fragkouli [12], Holger Fröhlich [13], Adrian Galiana-Bordera [14], Holger Hennig [15], Katja Herzog [15], Ida Holte Thorius [6], Bayrem Kaabachi [16], Jan Klein [17], Bogdan Kulynych [16], Levente Lippenszky [8], Luis Marti-Bonmati [14], Serena Elizabeth Marshall [2], Nagat Masued [7], Daniel Mensing [17], Matthias Müllenborn [6], Juan Parras Moral [1], Fotis Psomopoulos [12], Núria Queralt-Rosinach [18], Jean Louis Raisaro [16], Diego Valderrama [13], Hongxu Yang [8], Ahmed Youssef Ali [19], Santiago Zazo [1]

[1] Universidad Politécnica de Madrid (UPM), [2] DNV, [3] University of Oslo (UiO), [4] Helsinki University Hospital (HUS), [5] Università di Bologna (UNIBO), [6] Novo Nordisk, [7] Barcelona Supercomputing Center (BSC), [8] GE Healthcare, [9] TRAIN, [10] Humanitas Research Hospital, [11] Pfizer, [12] Centre for Research and Technology-Hellas (CERTH), [13] Fraunhofer SCAI, [14] Instituto de Investigación Sanitaria La Fe (IIS La Fe), [15] Fraunhofer ITMP, [16] Centre hospitalier universitaire vaudois (CHUV), [17] Fraunhofer MEVIS, [18] Leiden University Medical Center (LUMC), [19] Johnson & Johnson

Find Out More!

The IHI SYNTHIA project

Funded by the **Innovative Health Initiative (IHI)**, SYNTHIA is a public-private partnership advancing **synthetic data generation (SDG)** for biomedicine. It develops tools to tackle data scarcity and privacy, generating tabular, imaging, sequencing data and more across six disease areas: **lung cancer, breast cancer, multiple myeloma, DLBCL, Alzheimer's, and Type 2 diabetes** (*Fig. 1*). Outputs will be delivered via a **federated platform** to support research and innovation.



Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications.

This project is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101172872. The JU receives support from the European Union's Horizon Europe research and innovation programme, COCIR, EFPIA, Europa Bío, MedTech Europe, Vaccines Europe and DNV. The UK consortium partner, The National Institute for Health and Care Excellence (NICE) is supported by UKRI Grant 10132181.

Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

***Figure 1. Targeted Synthetic Data Applications**. SYNTHIA delivers purpose-built synthetic datasets and tools to meet specific clinical and research needs in real-world healthcare settings.*

A Comprehensive Review of SDG Methods in Biomedicine

As part of the SYNTHIA project, a **scoping review** was conducted to assess the state of the art, identify key challenges and opportunities, and define a shared understanding of SDG to guide future development. The review followed a structured methodology, focusing on high-impact publications and preprints from the past 5–10 years across databases including **PubMed, Scopus, Web of Science, IEEE Xplore,** and **arXiv**.

Overview of common SDG approaches

Approach	Strengths	Limitations
Statistical	Simple, interpretable	Privacy, scalability, limited dynamics
Machine learning	Flexible, powerful	Complex, compute-heavy, training risks
Simulations	Domain-grounded, efficient	Less adaptable, lower variability

Statistical methods

- **KDE:** Non-parametric, replicates distributions; limited privacy, high compute.
- **Gaussian Copulas:** Capture multivariate dependencies; struggle with scale/privacy.
- **Mixture Models (e.g., GMMs):** Model heterogeneity; newer versions handle mixed data.
- **Bayesian Networks:** Encode dependencies; some support private variants.
- **Oversampling (SMOTE/ADASYN):** Balance classes; simple but privacy-limited.

Machine learning methods

- **NODEs / ANODEs:** Model latent continuous-time dynamics; great for irregular time-series.
- **Neural Laplace**: Laplace-enhanced NODEs; better for sparse time points.
- **GNNs:** Learn on graph data; with variants for relational health data.
- **VAEs:** Probabilistic encoding; some handle missing/mixed data and temporal modeling.
- **GANs:** Realistic outputs via adversarial learning; unstable, mode collapse risk.
- **DDPMs:** Probabilistic noise-based generation; high-quality, but slow.
- **LLMs (e.g., GPTs):** Emerging for tabular/text; expressive but with privacy/ethics concerns.

Simulative methods

- **Imaging-Based:** Embed clinical features in clean images; efficient and realistic.
- **Physics/Chemistry Models:** Simulate interactions (e.g., contrast imaging).
- **Rule-Based Simulators (e.g., Synthea):** Generate synthetic EHRs from clinical rules.

Multimodal data

Multimodal synthetic data generation in healthcare leverages advanced AI techniques to integrate diverse data types, such as tabular, imaging, time-series, and omics, within unified frameworks. Methods include, transformer-based architectures, hybrid models, multimodal GANs, and diffusion models, which learn complex cross-modal dependencies.

Sequencing data

Synthetic sequencing data is generated using a variety of tools that range from basic short-read simulators to advanced models incorporating real sequencing error profiles and genomic features, enabling controlled benchmarking and testing of bioinformatics workflows. Recent developments include large language models like **Evo** [20], **GenSLMs** [21], and **xTrimoPGLM-100B** [22], which enhance sequence generation and functional prediction.

Spatial data

Spatial transcriptomics (ST) maps gene expression within tissue context, aiding cancer research and diagnostics. Synthetic ST uses computational models to generate realistic spatial datasets for method testing and data augmentation. Tools like **scDesign3** [23] help simulate tissue structure and gene activity for improved analysis development.

CONCLUSIONS

Ensuring data quality, clinical relevance, and minimizing bias requires robust **assessment**, **benchmarking** and **FAIR principles**. The evolving regulatory landscape (e.g., GDPR, HIPAA) calls for clear legal and ethical guidance. Future efforts must focus on **standardized evaluation metrics**, **stronger privacy protections**, and **expert-in-the-loop validation** to ensure real-world clinical utility.

SDG methods by data modality

Textual data

Synthetic medical text is generated using LLMs like GPT-4 and transformer-based frameworks such as **MedSyn** [1], which integrates Medical Knowledge Graphs, improving NLP tasks like ICD coding, NER, and de-identification. Tools like DeID-GPT automate removal of private information, enabling realistic, privacy-safe clinical text generation for research and clinical use.

Tabular data

Synthetic tabular data in healthcare is generated using classical statistical methods like GMMs, Copulas, and KDE, as well as machine learning techniques including decision trees (e.g., **Randomized Decision Trees** [2]), GANs (e.g., **MedGAN** [3], **CTGAN** [4]), VAEs (e.g., **TVAE** [4], **TabVAE** [5]), diffusion models (e.g., **TabDDPM** [6], **TabSyn** [7]), and LLMs (e.g., **GReaT** [8]). These methods address challenges like mixed data types, privacy, and complex dependencies, supporting applications from clinical trials to precision medicine.

Imaging data

Medical imaging data synthesis leverages tools like GANs, diffusion models, and hybrid approaches (e.g., **HA-GAN** [9]) to generate anatomically accurate and scalable synthetic images while addressing data scarcity and privacy concerns. Additionally, vision-language models such as **MedViLL** [10] and **Flamingo-CXR** [11] enable automated generation of clinically relevant radiology reports, improving workflow efficiency and diagnostic support.

Times series data

Synthetic signaling and time-series data generation utilizes advanced tools like **VAMBN** [12] and **VAMBN-MT** [13] for capturing complex temporal and multimodal dependencies, as well as **MultiNODEs** [14] for modeling continuous trajectories in latent space. For biomedical signals such as EEG and ECG, GAN variants (e.g., **CGANs** [15], **RGANs** [16]), **DDPMs** [17], **SynSigGAN** [18], and **DoppelGANger** [19] are employed to create realistic, high-fidelity synthetic data reflecting temporal dynamics and signal complexity.

[1] Kumichev G, Blinov P, Kuzkina Y, et al. *Lect Notes Comput Sci*. 2024;215–30.
[2] Vaidya J, Shafiq B, Asani M, et al. *AMIA Annu Symp Proc*. 2017;1695–704.
[3] Choi E, Biswal S, Malin B, et al. *arXiv*. 2017.
[4] Xu L, Skoularidou M, Cuesta-Infante A, et al. *arXiv*. 2019.
[5] Tazwar S, Knobbout M, Quesada E, et al. *Proc ICPRAM*. 2024.
[6] Kotelnikov A, Baranchuk D, Rubachev I, et al. *arXiv*. 2022.
[7] Zhang H, Zhang J, Srinivasan B, et al. *arXiv*. 2023.
[8] Borisov V, Seßler K, Leemann T, et al. *arXiv*. 2022.
[9] Sun L, Chen J, Xu Y, et al. *IEEE J Biomed Health Inform*. 2022;26:3966–75.
[10] Ye-Bin M, Hyeon-Woo N, Choi W, et al. *arXiv*. 2023.
[11] Tanno R, Barrett DGT, Sellergren A, et al. *Nat Med*. 2024.
[12] Gootjes-Dreesbach L, Sood M, Sahay A, et al. *Front Big Data*. 2020;3:16.
[13] Kühnel L, Schneider J, Perrar I, et al. *Sci Rep*. 2024;14:14412.
[14] Wendland P, Birkenbihl C, Gomez-Freixa M, et al. *NPJ Digit Med*. 2022;5:122.
[15] Wickramaratne SD, Parekh A. *Conf IEEE EMBC*. 2023;1–4.
[16] Esteban C, Hyland SL, Rätsch G. *arXiv*. 2017.
[17] Ho J, Jain A, Abbeel P. *arXiv*. 2020.
[18] Hazra D, Byun Y-C. *Biology (Basel)*. 2020;9.
[19] Lin Z, Jain A, Wang C, et al. *Proc ACM IMC*. 2020.
[20] Nguyen E, Poli M, Durrant MG, et al. *bioRxiv*. 2024.
[21] Zvyagin M, Brace A, Hippe K, et al. *bioRxiv*. 2022.
[22] Chen B, Cheng X, Li P, et al. *bioRxiv*. 2023.
[23] Song D, Wang Q, Yan G, et al. *Nat Biotechnol*. 2024;42:247–52.