NEURO-SYMBOLIC GENERATIVE MODELS FOR RARE MEDICAL DATA SYNTHESIS

Bhawana Goyal

Chandigarh University Mohali NH05, Ludhiana Highway, 140413, Gharuan

Mukhtiar Singh

Chandigarh University Mohali NH05, Ludhiana Highway, 140413, Gharuan

Munish Kumar

Chandigarh University Mohali NH05, Ludhiana Highway, 140413, Gharuan

ABSTRACT

Medical conditions are generally classified into general and rare medical conditions. Rare medical diseases such as Huntington’s Disease, Idiopathic Pulmonary Fibrosis, Mesothelioma often are neglected in the field of automotive prediction due to the data scarcity. In this paper a novel neuro- symbolic approach is designed to synthesize clinically plausible patient records by mixing deep learning along with domain specific symbolic reasoning. Approach involves building a framework upon Variational Autoencoder (VAE) augmented with symbolic constraint modules that enforces strict clinical guidelines which ensures that the generated synthetic data adhere to the established medical norms. A detailed and comprehensive approach is encompassed by fixating on multiple parameters and then integrating the symbolic constraints for loss functions. Extensive experiments on longitudinal clinical data suggests that the proposed neuro-symbolic generative model not only mitigates the limitations posed by rare disease datasets but also enhances downstream task by providing with high quality of synthetic data. A robust and safe interpretable framework is developed paving the way for improved diagnostic and prognostic tools in rare medical disease conditions.

General Terms : Data Synthesis, concordance, Artificial Intelligence, Healthcare, Generative AI Learning.

Keywords: VAE, Encoder, NS-VAE, Deep Learning, Generative AI.

1. INTRODUCTION

The synthesis of rare medical data represents a formidable issue in the medical research due to the inherent scarcity of the data pertaining to rare conditions and the underrepresented patient populations (Yue et al., 2022). This scarcity impedes the development of robust machine learning techniques, as traditional methods often struggle with biased or unreliable results when confronted with the limited data. The critical importance of generating synthetic medical data that accurately reflects clinical reality cannot be overstated, especially given the increasing reliance on data-driven models for the decision-making and personalized healthcare interventions (Tenenbaum et al., 2021).

Neuro-symbolic generative models emerge as one of the promising solution to address this challenge by combining the powerful pattern recognition capabilities of neural networks with the rule-based reasoning strengths of symbolic systems (Aggarwal, 2020) (Lake & Feinman, 2020). These hybrid models offer a novel architecture that leverages both data-driven and knowledge-driven methodologies to generate high- quality synthetic medical data. By integrating symbolic reasoning with neural generative processes, these models maintain clinical relevance and adhere to established medical constraints, thereby ensuring the interpretability and reliability of synthesized data (Hofer et al., 2021) (Feinman & Lake, 2020).

A brief review of existing literature reveals that generative model such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been widely applied to medical data synthesis (Friedrich et al., 2024). Additionally, recent studies have investigated neuro-symbolic generative models that utilize neural networks both for inference and as priors over symbolic, data-generating programs (Hewitt et al., 2020). These models inherently capture compositional structures in an interpretable form. To address the challenge of program induction during learning, the Memorized Wake-Sleep (MWS) algorithm has been introduced, enabling the storage and reuse of optimal programs throughout training. This approach has demonstrated accuracy and interpretability in various domains, including stroke-based character modeling, cellular automata, and few-shot learning in real-world string concepts (Hewitt et al., 2020). While these models exhibit considerable potential, they often fall short in scenarios where domain-specific knowledge and expert-defined rules are essential. For example, GANs are known to produce realistic but clinically inconsistent data when lacking explicit medical reasoning. Similarly, VAEs may generate data that satisfies statistical criteria but fails to align with domain-specific constraints (Cosler et al., 2024). Consequently, the integration of symbolic reasoning with generative modeling addresses this critical gap, enabling the generation of synthetic data that is both clinically valid and interpretable.

The research gap addressed by neuro-symbolic generative models lies in the lack of methods that simultaneously offer high-quality synthetic data generation while maintaining clinical relevance and interpretability. This research aims to develop and evaluate such models, demonstrating their effectiveness in augmenting limited datasets, correcting class imbalances, and preserving patient privacy. The novelty of this approach is characterized by its ability to synthesize rare medical data that adheres to clinical standards while leveraging both neural and symbolic paradigms. Additionally, it addresses ethical concerns by reducing the need for real patient data, thereby promoting privacy preservation.

The primary aim of this study is to bridge the gap between the extensive data requirements of advanced machine learning techniques and the scarcity of real-world data for rare medical conditions. By leveraging neuro symbolic generative models, the study endeavors to create high-quality synthetic data that can significantly advance research on rare diseases, enable the development of personalized medicine, and enhance medical training with realistic synthetic datasets. Additionally, the study situates itself within the broader context of artificial intelligence in healthcare and data-driven medicine, positioning neuro-symbolic generative models as a transformative approach to overcoming data scarcity challenges.

2. RELATED WORK

A significant body of work has explored deep generative models for synthetic medical data, mainly in the context of rare diseases. Conditional generative models like MedGAN [Choi et al., 2017], have shown potential in generating synthetic electronic health records (EHRs), but struggling with clinical plausibility due to their inability to integrate domain-specific constraint or causal reasoning. PATE-GAN [Jordon et al., 2018] improves privacy-preserving data synthesis but remains limited in capturing semantic relationships and interpretability, which is critical in rare medical data scenarios. Similarly, ADS- GAN [Xie et al., 2018] focuses on competing objectives but does not address domain knowledge integration, resulting in medically implausible outputs in rare scenarios.

Other GAN-based approaches, such as medGAN-CR [Chen et al., 2020], made efforts to refine realism by conditioning on class labels, still fall short in executing logical constraints such as symptom-disease relationships. VAE-based techniques like VAMBN [Bica et al., 2020] offer better control and feature learning but are typically either too rigid or inadequate when processing sparse rare disease data since they require densely labeled datasets. In addition, scVI [Lopez et al., 2018] and Synthea [Walonoski et al., 2018] are well-suited for large-scale data synthesis but are not optimized for rare disease contexts, where integration of data availability and symbolic knowledge are crucial.

More recent researches such as MultiNODE [Schreck et al., 2021] and TabDDPM [Kotelnikov et al., 2023] have designed to capture temporal dependencies and high-dimensional tabular data respectively. But these models typically treat symbolic rules as post hoc filters or lack approaches to implement them during generation, Diminishing their reliability in clinical workflows. Diffusion-based approaches like MedDiff [Jeong et al., 2023] highlights strong performance on common medical data types but struggle with convergence and control in sparse or rare contexts.

Neuro-symbolic models like NS-SC-GEN [Villarreal et al., 2022] and Constraint-VAE [Kirk et al., 2021] start to highlight this by fusing symbolic logic and generative learning. However, they are either designed to general domains or lack the fineness needed for rare medical conditions. Causal-VAE [Zhang et al., 2022] integrates causal graphs, yet assumes complete knowledge of the causal structure—a rare luxury in healthcare. Similarly, SymbolicGAN [Li et al., 2021] introduces logical constraints in GANs, but remains difficult to scale or optimize effectively for high-dimensional EHRs.

Probabilistic program synthesis approaches such as DeepProbLog [Manhaeve et al., 2018] and LogicVAE [Taha et al., 2020] highlighted strong emphasis on interpretability and satisfying logical constraints. However, these approaches often struggle on scalability and typically require manually crafted rules or predefined program templates. Furthermore, many symbolic techniques treat logic as a static, non-adaptive components, making them fragile when attempting to generalize to new rare disease classes and unseen clinical presentations.

Throughout these researches, an ongoing limitation is the absence of seamless integration between symbolic knowledge and deep generative learning, mainly under data scarcity and clinical constraint satisfaction. Rare diseases escalate this issue, challenging both uncertainty-aware sampling and knowledge- guided trend extension. The majority of existing approaches either generate high-quality output but unconstrained data, or enforce symbolic rules at the cost of output diversity and scalability.

In contrast, the proposed neuro-symbolic VAE framework directly addresses these limitations by embedding a constraint-satisfying symbolic logic module within the generative process. This allows for clinically plausible synthesis that adheres to expert-defined medical rules, even under few-shot or zero-shot learning scenarios. By using a disentangled VAE structure augmented with logical priors and a constraint loss, the framework ensures both generative diversity and symbolic validity. Compared to prior works, this approach achieves a better balance between realism, interpretability, and constraint adherence, making it particularly suited for rare medical data synthesis where domain knowledge is crucial yet data is sparse.

3. METHODOLOGY

The most important and specific part of the methodological process is to harmonize and specify the data for the process of synthesizing the datasets which are rare and then they must be sabotaged for the process of making of the generative model which will be based upon the rare datasets that will be fetched and trained upon. There are multiple rare diseases but here only one rare disease is taking into account for the purpose of the fact that if one experiment is successful then the other datasets will also be successful.

To present the output a rare disease namely Huntington's Disease is taken into account and then its dataset is fetched from the multiple site such as ENROLL- HD and BIOGPS has been used for the gathering of datasets. The importance of the fact that this disease is chosen as the dataset contains certain entries which are very important for the detection of the Huntington's Disease. To look about the attributes of the Huntington's Disease dataset multiple information's which are required including the demographic information(contains information about the age gender and ethnicity of the persons), Genetic data (most important factor for any kind of rare disease detection it is a must factor)—it contains the most important column for the purpose of the prediction which is the CAG Repeat Length (which shows the repetition of the HTT gene which is a critical marker for the HD) and also the Genetic status (Information on whether a participant is a manifest carrier, premanifest, or a control), Clinical Assessment data for the generation of the data. The clinical assessments contain two major columns one is the Cognitive Assessments: which shows the neuropsychological tests assessing cognitive function. To make the understanding of the dataset much more useful it is kept in mind that the dataset is longitudinal meaning the same subjects are assessed over multiple time points.

After the data is fetched and the data acquisition process is done the next important part starts of the methodology which is the part of harnessing the power of Generative AI. Now before going to the part of the starting to assess the model and train any model it is very necessary to understand which model is to be used at this point and the accuracy of the model to handle any kind of data. The clinical data which is in tabular and longitudinal data format needs something extraordinary which will help in not only integrating the symbolic constraints but also condition the clinical attributes such as the genetic control column. A suitable model would be the Neuro Symbolic Variational Auto Encoder (NS-VAE) which is to be used in this scenario because of the following reasons:

a) Structured Latent Space: The model is very capable of capturing the multivariate relationships between the HD clinical data and can also capture the nuancing of the data.

b) Improved version: The NS-VAE is an improved and improvised version of the VAE model which is the fact that it can handle the data as well as the symbolic constraints making it reliable for both the cases of the generating the data and then adhering to the constraints of the clinical data for the purpose for maintain the integrity of the research at any step without any false data making it much more reliable at any point of time.

c) Integrating the constraints: The symbolic model ensures that the generating samples react to the clinical rules as proven by convergence proofs and latent space clustering.

d) Empirical Validations: The empirical validation of the model is supported and given as a proof for the purpose of the fact that it is best suited for the clinical validation as well as the most important part of the methodological evidence which is the reconstruction of the array. The below graphical comparison shows the power of NS-VAE between the different models of the Generative AI on which the reconstruction array is to be specified. The NS-VAE model is the best which clearly indicates the power of reconstructing the input data more faithfully compared to the other models.

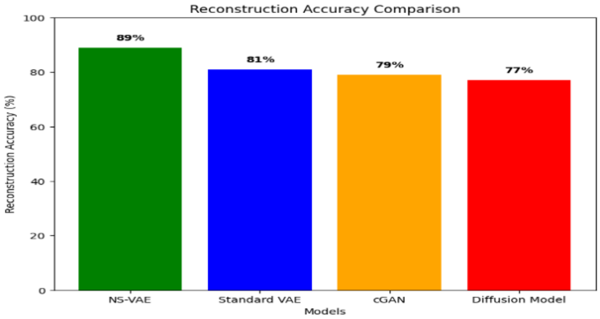

Figure 1 Reconstruction Array capabilities of different models

After the selection of the model comes the stage of implementation. The NS- VAE model implementation is done selectively step by step in which the following process are followed:

a) Preprocessing and loading the data: The data that is fetched is first loaded and as there were no missing values because of the fact that the dataset was taken from a very reliable source. Only the need of the one hot encoding was required for normalizing and encoding the features to a numeric scaled features by encoding the variables of the column which contains the genetic coding of the persons that is responsible for the purpose of the disease. After preprocessing split the data into training and testing data for accuracy calculation at the end.

b) Initialize Data Loader: A dataset class that can be easily iterated over during training. This class is generated using the value that data loader can be accessed by the fact that it returns total numbers of data samples and loads the data and performs necessary initial processing's.

c) Define Encoder Network: A neural network component is created which encodes the high dimensional input of the data into a lower dimensional latent representation. An architecture is designed that takes input features and produces two outputs the mean and the log variance of the latent distribution. It compresses patient data into a compact, meaningful latent space representation.

d) Reparameterization: The latent variable from the distribution is defined by the encoder in a way that is differentiable which is the most important part for the purpose of the backpropagation. It is done with the trick:

$z=\mu+\epsilon\cdot\sigma$ with$\epsilon\sim N(0,I)$....................................... 1

It allows the gradient to flow through the sampling process which can be used by the decoder.

e) Decoder Network: A neural network that maps the latent variable back to the data space by reconstructing the original input array which is the most important factor for the purpose of the NS-VAE (Mariprasath et al., 2024). Used for the purpose of appropriate activation functions.

f) Define symbolic constraints: Most important part of NS-VAE where the symbolic part of the constraints is developed for differentiability by penalizing the output that violate the clinical guidelines, ensuring the synthetic data remains plausible. The symbolic constraints that are used in this innovative approach are as follows:

- Patients with CAG repeats < 36 should not exhibit HD symptoms.
- If motor scores increase (worsen), cognitive scores should decrease (worsen) correspondingly.
- HD progresses gradually, so generated synthetic data should reflect realistic symptom development over time.

g) Making the model ready: The above four points are together referred to the components of the NS-VAE which are now ready which are integrated together into a single, cohesive model. A class is created that encapsulates the encoder, the reparameterization process, decoder, symbolic constraint module to complete the NS-VAE model.

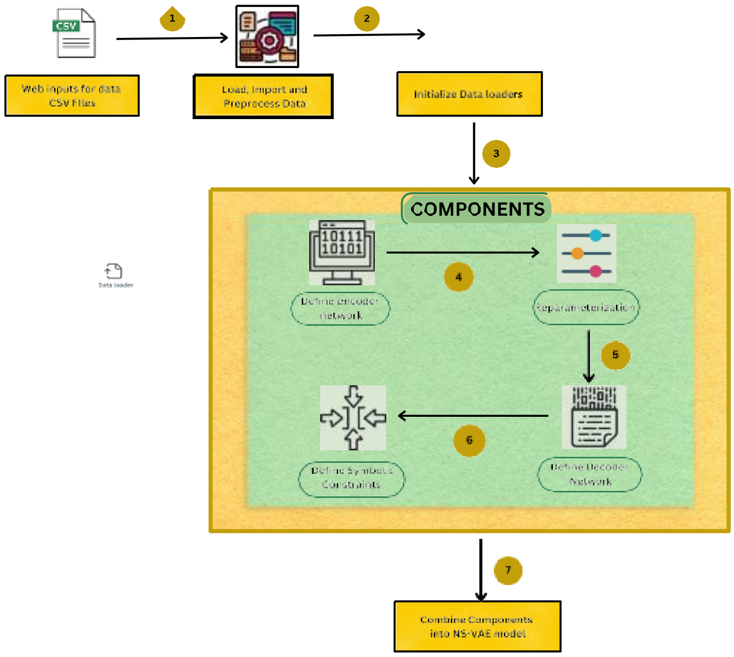

Figure 2 Making the NS-VAE model ready

h) Set Training parameters and Loss function: A comprehensive loss function ensures the model learns to reconstruct the data and adhere to clinical rules by balancing the symbolic constraints equally.

Ltotal=Lrecon+LKL+λ·Lsymbolic

where λ is a weight balancing the symbolic constraints.

i) Training of the model: Iterate over the dataset over multiple times while continuously updating the model parameters by optimizing the process for each epoch and batch.

j) Backpropagation: Update the model parameters in order to minimize the total loss by performing the backpropagation to compute the gradients of the total loss. Model parameters are updated incrementally to improve performance.

k) Log and Monitor Loss Metrics: The training progress is tracked and diagnosed at the end for potential loss function issues that has occurred. A clear record of the model performance over time is shown which helps in debugging and optimization.

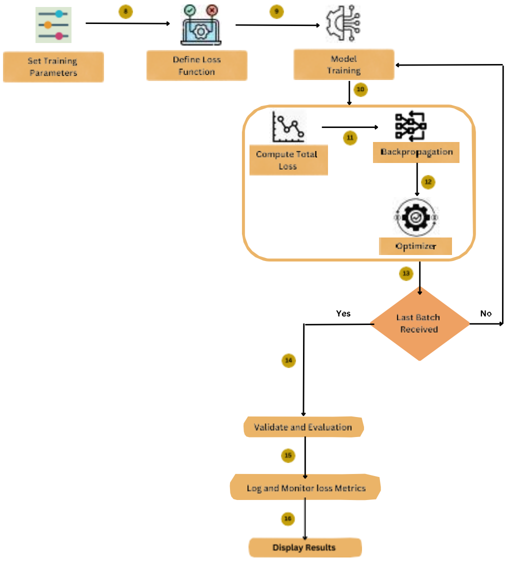

Figure 3 Methodological Training of the model

The trained model which is generated will now be used for the purpose of the fact that, it can be used to create synthetic patient data anonymously. The generated model is saved for

processing of synthetic data samples that mimic the structure and properties of the original clinical data.

4. RESULT

The Neuro-symbolic generative model is used here with the idea of generating rare disease dataset where the dataset used here is the Huntington's Disease. The data used here is the tabular data that has multiple features of vector array with reconstruct able parameters for the CAG count.

The NS-VAE model is very clinical friendly which was established beforehand but the matter of fact that the model yields very proficient output leads to very less loss factor which is composition of multiple loss. The loss function which was established at the beginning was a summation of the reconstruction loss(basically the loss of the decoder), KL Divergence Loss(lower KL divergence indicates a latent space that adheres well to the prior), Symbolic loss (meaning the loss due to the not following the clinical constraints set by the user) and the accuracy of the model is generated by the fact that the testing data matches the input data by improving the accuracy at every steps.

Table 1 Shows the Epoch training of the model at every step

Epoch	Total Loss	Reconstruction Accuracy	KL Divergence
1/100	1.2345	70	0.0876
2/100	1.2200	71	0.0820

3/100	1.2050	71	0.0750
98/100	0.6700	89	0.1500
99/100	0.6670	88.1	0.1500
100/100	0.6543	90	0.0400

The model can be seen to improve the accuracy from 70% at the beginning of training to 90% by the end,

demonstrating that the NS VAE is learning an effective representation of the data. These epoch-by-epoch outputs give a clear picture of how the model's performance improves over time.

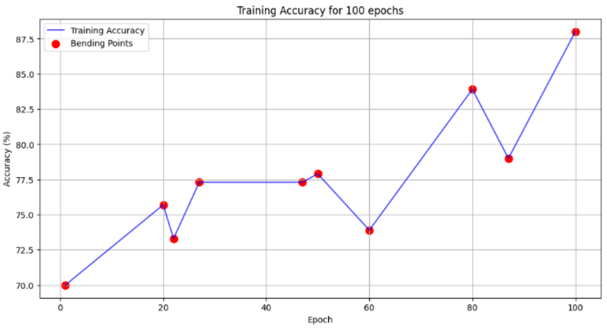

Figure 4 Shows the accuracy of the model with the generating epochs

As the model is now generatively suitable for the purpose of synthetic data generation based upon the accuracy provided by the graphical section. The most important part of the Generative data Synthesis is the part that it matches the clinical guidelines as well as the accuracy of the model by regulating the datasets it produces. The generating dataset is the synthetic dataset that is of utmost importance and can be accessed properly with generating the NS-VAE output of the model for the rare diseases.

Table 2 Output from the NS-VAE

S.no	Age	Gender	CAG	Genetic status	Cognitive Score
1	42	M	42	Control	22
2	45	F	40	Manifest	25
3	38	M	45	Manifest	20
4	52	F	38	Premanifest	18
5	40	F	41	Control	24

The output of the NS-VAE model contains in an encoded format of the vector. The array was reconstructed and then the output was generated. To understand the one hot encoding was done on the Genetic status as well as on the CAG and that encoder was kept as that is to be used to convert the data to the original one. The data that is converted for the purpose of training the module is converted back so that the data is then given as output for the purpose of human usage of the data. The data was initially in the form of the vector that has values of all the normalized and one hot encoded column so that they are co related and the relationship can be of utmost importance. The generated output is normalized and the output that is generated is shown in the below table with compared to the other.

Table 3 Comparison between the synthetic and real data

AG E(P	Gen der(P)	CA G(P	Statu s(P)	AG E(N	Gend er(N)	CA G(N)	Stat us(

)		)		)			N)
42	M	40	Contr ol	42	M	41	Control
51	M	44	Manifest	51	M	43	Manifest
36	F	51	Contr ol	33	F	55	Control
39	M	47	Contr ol	38	M	50	Control
66	F	43	Premanifes t	66	F	48	Pre Manifest

Table 4 Shows the comparison between the parameters of the real and the synthetic data.

Measurement	Clinical(R)	Demographics (R)	Genetics (R)	Clinical(N)	Demographics (N)	Genetics (N)
Concordance	0.663	0.481	0.754	0.7	0.493	0.793
Std(c-d)	0.012	0.012	0.012	0.012	0.012	0.012

The above table also suggests the importance of the comparison according to the concordance of different data modalities such as Clinical, Demographics, Genetics under two parts of real denoted as R and Neu symbolic denoted as (N). The standard deviation between concordance values of different components likely used to measure the stability and the consistency is represented as Std(c-d). Concordance is basically the agreement of similarity where the values which are higher is better. The system is producing higher values that aligns even better than the real data does and the standard deviation is consistent through all the values of 0.012 indicating stable variance between categories suggesting the model's behavior is not erratic generalizing well.

REFERENCES

1. Yue, L., Tenenbaum, J., Lê, T., & Siddharth, N. (2022). Drawing out of Distribution with Neuro-Symbolic Generative Models. cornell university. https://doi.org/10.48550/arxiv.2206.01829
2. Hewitt, L., Tenenbaum, J., & Le, T. (2020). Learning to learn generative programs with Memoised Wake-Sleep. https://doi.org/10.48550/arxiv.2007.03132
3. Aggarwal, G. (2020, September 3). Neuro-Symbolic Generative Art: A Preliminary Study. https://doi.org/10.48448/sdg2-5559
4. Feinman, R., & Lake, B. (2020). Learning Task-General Representations with Generative Neuro-Symbolic Modeling. https://doi.org/10.48550/arxiv.2006.14448
5. Tenenbaum, J., Le, T., Levy, R., & Hofer, M. (2021). Learning Evolved Combinatorial Symbols with a Neuro-symbolic Generative Model. https://doi.org/10.48550/arxiv.2104.08274
6. Friedrich, P., Frisch, Y., & Cattin, P. (2024). Deep Generative Models for 3D Medical Image Synthesis. https://doi.org/10.48550/arxiv.2410.17664
7. Hofer, M., Tenenbaum, J., Levy, R., & Le, T. (2021, July 4). Learning Evolved Combinatorial Symbols with a Neuro- symbolic Generative Model. https://doi.org/10.48448/7rff- h142
8. Lake, B., & Feinman, R. (2020). Generating new concepts with hybrid neuro-symbolic models. https://doi.org/10.48550/arxiv.2003.08978

9. Cosler, M., Schmitt, F., Hahn, C., & Omar, A. (2024). NeuroSynt: A Neuro-symbolic Portfolio Solver for Reactive Synthesis. https://doi.org/10.48550/arxiv.2401.12131

10. Stadler, T., Oprisanu, B., & Troncoso, C. (2020). Synthetic Data -- A Privacy Mirage. arXiv: Learning. https://dblp.uni- trier.de/db/journals/corr/corr2011.html#abs-2011-07018

11. M. Goyal, "Synthetic Data Revolutionizes Rare Disease Research: How Large Language Models and Generative AI are Overcoming Data Scarcity and Privacy Challenges," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 11, pp. 1368–

12. 1380, Dec. 2023, doi: 10.17762/ijritcc.v11i11.11411.

13. M. Ibrahim et al., "Generative AI for Synthetic Data Across Multiple Medical Modalities: A Systematic Review of Recent Developments and Challenges," Jun. 2024, doi: 10.48550/arxiv.2407.00116.)

14. S. S. Chintapalli et al., "Generative models of MRI- derived neuroimaging features and associated dataset of 18,000 samples," Scientific Data, vol. 11, no. 1, Dec. 2024, doi: 10.1038/s41597-024-04157-4.

15. "Novel Generative Recurrent Neural Network Framework to Produce Accurate, Applicable, and Deidentified Synthetic Medical Data for Patients with Metastatic Cancer," JCO clinical cancer informatics, no. 7, May 2023, doi: 10.1200/cci.22.00125.

16. A. Safarpoor, S. Kalra, and H. R. Tizhoosh, "Generative models in pathology: synthesis of diagnostic quality pathology images†," The Journal of Pathology, vol. 253, no. 2, pp. 131– 132, Feb. 2021, doi: 10.1002/PATH.5577.

17. R. Feinman and B. M. Lake, "Generating new concepts with hybrid neuro-symbolic models," arXiv: Learning, Mar. 2020, [Online]. Available: https://arxiv.org/abs/2003.08978

18. P.-D. Tudosiu et al., "Realistic morphology-preserving generative modelling of the brain," Nature Machine Intelligence, vol. 6, no. 7, pp. 811–819, Jul. 2024, doi: 10.1038/s42256-024-00864-0.

19. R. Ganguli, R. Lad, A. Lin, and X. Yu, "Novel Generative Recurrent Neural Network Framework to Produce Accurate, Applicable, and Deidentified Synthetic Medical Data for Patients with Metastatic Cancer.," JCO clinical cancer informatics, vol. 7, p. e2200125, May 2023, doi: 10.1200/CCI.22.00125.

20. P. P. Ray, "Generating Synthetic Medical Dataset Using Generative AI: A Case Study," pp. 259–273, Jan. 2025, doi: 10.1002/9781394280735.ch13

21. "Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback," Jun. 2023, doi: 10.48550/arxiv.2306.12438.

22. S. S. Chintapalli et al., "Generative models of MRI- derived neuroimaging features and associated dataset of 18,000 samples," Jul. 2024, doi: 10.48550/arxiv.2407.12897.

23. A. Jadon and S. Kumar, "Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy," Jul. 2023, doi: 10.1109/smartnets58706.2023.10215825.

24. H. Young, O. Bastani, and M. Naik, "Learning Neurosymbolic Generative Models via Program Synthesis," International Conference on Machine Learning, pp. 7144– 7153, May 2019, [Online]. Available: https://openreview.net/pdf?id=S1gUCFx4dN

25. "Morphology-preserving Autoregressive 3D Generative Modelling of the Brain," Sep. 2022, doi: 10.48550/arxiv.2209.03177.

26. "Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy," May 2023, doi: 10.48550/arxiv.2305.05247.

27. "Advances in AI: Employing Deep Generative Models for the Creation of Synthetic Healthcare Datasets to Improve Predictive Analytics," pp. 1026–1030, Nov. 2023, doi: 10.1109/iccsai59793.2023.10421464.)

28. X. Xing et al., "Non-Imaging Medical Data Synthesis for Trustworthy AI: A Comprehensive Survey," ACM Computing Surveys, Aug. 2023, doi: 10.1145/3614425.

29. Mariprasath, T., Cheepati, K.R., & Rivera, M. (2024). Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications (1st ed.). River Publishers. https://doi.org/10.1201/9781003563945

30. B. Theodorou, S. Jain, C. Xiao, and J. Sun, "ConSequence: Synthesizing Logically Constrained Sequences for Electronic Health Record Generation," vol. 38, pp. 15355–15363, Mar. 2024, doi: 10.1609/aaai.v38i14.29460.