Advancing Cancer Research With Synthetic Data Generation in Low-Data Scenarios

Patricia A. Apellániz, Borja Arroyo Galende, Ana Jiménez, Juan Parras, and Santiago Zazo

***Abstract*—The scarcity of medical data, particularly in Survival Analysis (SA) for cancer-related diseases, challenges data-driven healthcare research. While Synthetic Tabular Data Generation (STDG) models have been proposed to address this issue, most rely on datasets with abundant samples, which do not reflect real-world limitations. We suggest using an STDG approach that leverages transfer learning and meta-learning techniques to create an artificial inductive bias, guiding generative models trained on limited samples. Experiments on classification datasets across varying sample sizes validated the method's robustness, with further clinical utility assessment on cancer-related SA data. While divergence-based similarity validation proved effective in capturing improvements in generation quality, clinical utility validation showed limited sensitivity to sample size, highlighting its shortcomings. In SA experiments, we observed that altering the task can reveal if relationships among variables are accurately generated, with most cases benefiting from the proposed methodology. Our findings confirm the method's ability to generate high-quality synthetic data under constrained conditions. We emphasize the need to complement utility-based validation with similarity metrics, particularly in low-data settings, to assess STDG performance reliably.**

***Index Terms*—Generative models, inductive bias, medical data scarcity, synthetic data generation and tabular data.**

I. Introduction

THE revolutionary potential of Deep Learning (DL) in healthcare is undeniable, with significant advancements in medical image analysis, disease diagnosis, and personalized treatments [23], [35], [56]. DL-powered tools analyze extensive patient data to identify patterns and risk factors, fostering innovative treatments and expanding healthcare access through telemedicine, particularly in underserved areas. However, DL in healthcare faces substantial challenges, primarily due to the scarcity of high-quality medical data [41], [52], [56]. Strict privacy regulations, patient confidentiality, and data heterogeneity across institutions hinder data acquisition and sharing. These issues are particularly pronounced in rare diseases, where limited patient data compounds the challenge. Even when large datasets exist, accessibility is restricted by privacy concerns and legal barriers, delaying research progress and clinical translation [11], [32]. Synthetic Data Generation (SDG) has emerged as a promising solution to overcome data limitations, particularly in medical artificial intelligence [38]. Advances in generative DL models, such as Stable Diffusion [46], DALL-E [45], and large language models like GPT [14] and ChatGPT [39], have revolutionized data generation in images, text, and video. These advancements demonstrate the potential of SDG in creating realistic and high-quality synthetic datasets, which could accelerate medical research.

In healthcare, SDG addresses privacy concerns by generating entirely new synthetic datasets that mimic the statistical properties of real data while ensuring anonymity [8]. Unlike traditional anonymization methods, which cannot guarantee complete privacy [28], [44], SDG enhances dataset diversity and model robustness. By replicating the distribution and correlations of original datasets, SDG provides realistic data that are critical for research, especially in contexts where real datasets are unavailable or restricted.

Research in SDG has seen a notable increase, with expectations for broader adoption in the future [16]. There is a pressing need to consolidate this body of knowledge to enhance comprehension and application. Current methods often rely on domain-specific data structures, limiting their generalizability. This limitation is quite evident in the medical field, where most SDG models focus on medical imaging [30]. At the same time, tabular data (EHRs, clinical trials, and laboratory datasets) pose unique challenges due to their complexity and diverse data types (categorical, ordinal, dates, etc.). Tabular data offer flexibility and broad applicability across various domains. These datasets often contain interconnected features, including explicit identifiers, quasi-identifiers, and sensitive attributes, needing robust SDG models capable of handling their intricate and longitudinal nature.

The need to generate synthetic medical data has led to the exploration of different generative techniques, ranging from classical oversampling methods to more advanced Deep Generative Models (DGMs). Traditional approaches, such as the Synthetic Minority Oversampling Technique (SMOTE) [15] and ADASYN [24], have been widely employed to address class imbalance, particularly in cancer-related datasets by interpolating new samples from existing minority class examples [2],

Received 2 January 2025; revised 23 April 2025 and 25 June 2025; accepted 30 July 2025. Date of publication 4 August 2025; date of current version 4 February 2026. This work was supported by European Union Horizon 2020 Research and Innovation Program through Projects GenoMed4All and SYNTHEMA under Grant 101017549 and Grant 101095530. *(Corresponding author: Patricia A. Apellániz.)*

The authors are with Information Processing and Telecommunications Center, ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: patricia.alonsod@upm.es).

Digital Object Identifier 10.1109/JBHI.2025.3595371

[37]. Although these methods improve model generalization, they do not create entirely independent synthetic datasets but rather variations of the original data points. More advanced hybrid models, such as Autoencoder-Generative Adversarial Network architectures [1], [19], combine feature extraction capabilities with generative modeling to enhance data diversity. Although these techniques are not designed explicitly for SDG, they demonstrate the increasing role of generative modeling in addressing data limitations in medical research.

Prominent SDG models for tabular data include those based on Generative Adversarial Networks (GANs), such as MedGAN [17], DCGAN [42], and CTGAN [55], and Variational Autoencoder (VAE)-based models like TVAE [55], VAEM [34], and the VAE with a Bayesian Gaussian Mixture (BGM) [5]. These models have demonstrated strong performance in generating realistic synthetic tabular data but rely on large datasets comprising thousands of samples, limiting their applicability in medical contexts where data scarcity is common. For example, MedGAN and CTGAN are typically evaluated on datasets with 20 k–500 k samples [17], [55], which is unrealistic for many healthcare applications. Consequently, advancing STDG models that can accommodate the unique challenges posed by medical tabular data, including its scarcity and complexity, remains a critical area of research.

In addressing this challenge, [7] proposes a novel methodology for generating realistic synthetic tabular data using DGMs in data-scarce environments. This approach introduces artificial inductive bias through transfer learning and meta-learning techniques, enhancing the quality of SDG. The methodology has been validated using CTGAN [55] and a VAE with a Bayesian Gaussian Mixture model [5]. Findings demonstrate that incorporating artificial inductive bias improves the realism and reliability of synthetic data, addressing key limitations in current SDG techniques. The authors also emphasize the importance of robust validation techniques for SDG models. They propose using divergences, as outlined in [4], to evaluate the similarity between real and synthetic data while considering both marginal and joint distributions. This approach ensures that synthetic data are statistically similar to real data and maintains correlations critical for practical applications.

Given the natural data scarcity in the medical field, the proposed methodology of [7] suggests extending this approach to healthcare data. Therefore, in this work, we propose to explore further the potential of this methodology to address the challenge of STDG in healthcare. We aim to demonstrate this methodology's effectiveness in a data scarcity setting. Our contribution can be justified as follows.

- *Addressing Data Scarcity in Medical Tabular Data:* This study targets small, heterogeneous tabular datasets commonly encountered in healthcare, such as patient records, clinical trials, and epidemiological studies, which are areas not extensively covered in the current state of the art. By generating synthetic data for rare diseases and other data-scarce scenarios, the proposed approach enables robust analyses that the lack of real data would otherwise limit.
- *Diverse Medical Data Types and Tasks:* The methodology is evaluated across diverse datasets, including Survival Analysis (SA) and classification tasks, to assess its applicability to realistic medical scenarios. Cancer-related SA datasets are emphasized due to their scarcity and complexity, while larger classification datasets are used to establish baseline performance. This comparison demonstrates the methodology's ability to enhance performance in worst-case scenarios and achieve optimal results when sufficient data are available.
- *Introduction of Robust Validation Techniques:* This work incorporates similarity validation using divergences [4] and clinical utility validation to ensure that synthetic data are statistically accurate and practically useful. By analyzing the relation between these validation criteria, the study offers deeper insights into the effectiveness of synthetic data.
- *Extension to General Use:* The proposed methodology not only supports task-specific validation but also facilitates general-purpose data analysis and visualization. The generated data could be publicly available, promoting collaboration and open science in medical research.

The paper is structured as follows: The Introduction outlines the significance of STDG in the medical domain and the challenges of data scarcity. The Methods section details the methodology implementation from [7], specifying models and validation techniques. The Results section describes the datasets (SA and classification), experimental settings, and findings and demonstrates the proposed methodology's effectiveness in generating synthetic data that resembles and functions like real data. The Conclusion highlights key contributions and implications for the medical field and proposes future research directions, including creating open-source databases.

II. Method

This section develops from the STDG framework previously proposed in [7], extending it to the medical domain. Before detailing the methodology and validation strategy, we summarize the key motivations and contributions of this work.

While the STDG methodology introduced in [7] demonstrated effectiveness in general tabular data, medical datasets present additional challenges that significantly hinder generative performance. These include high-dimensional and sparse data, missing values, strong inter-feature correlations, heterogeneity across patients, and label noise—all exacerbated by limited sample sizes and privacy restrictions [26], [50]. Medical tabular datasets are particularly complex due to their multimodal nature, irregular temporal structure, unbalanced feature distributions, and the presence of multiple tables with relational dependencies [36], [53].

Methods successful in other domains cannot be assumed to generalize to medical data without adaptation and validation. In fact, a growing body of research has shown that deep learning models require domain-specific adjustments to handle the unique characteristics of healthcare data. In the case of medical tabular data, standard deep neural networks often underperform compared to tree-based models unless specifically designed or trained to account for missingness patterns, categorical

heterogeneity, or relational structure [9]. Similar observations apply to other modalities in healthcare: for example, in medical imaging, convolutional neural networks need to be adapted to handle domain shifts, annotation scarcity, and class imbalance [57]; in clinical text mining, language models must be pre-trained or fine-tuned on domain-specific corpora such as EHR notes to reach acceptable performance [33]. These findings consistently reinforce the notion that adapting deep learning models to the medical domain is not only beneficial but often essential for achieving reliable and clinically relevant outcomes.

In this work, we extend and apply the STDG methodology to this highly complex domain, introducing three key contributions: (1) a full adaptation of the STDG strategy to health-related tabular data, including cancer SA datasets; (2) the design of an extensive medically grounded validation protocol, which is complementary to the original one, combining both similarity and clinical utility metrics across multiple scenarios; and (3) empirical evidence that supports the feasibility and robustness of STDG under these medical data regimes.

A. STDG Methodology

Building upon the methodology proposed by [7] on artificial inductive bias in STDG, our study applies this framework specifically to medical research. This approach introduces several contributions suited to the medical domain, particularly its effectiveness in scenarios with limited data availability, a common constraint in medical research due to data scarcity, time constraints, costs, and privacy concerns. The original work did not explore this methodology with medical datasets, prompting our investigation.

The original methodology introduces a novel approach to address the challenge of generating high-quality synthetic tabular data when faced with limited real data. The core concept incorporates an artificial inductive bias, a set of inherent preferences or assumptions integrated into the learning model during training. In scenarios where extensive training data (Big Data) are available, DGMs effectively capture the underlying data distribution, facilitating the generation of realistic synthetic data. In contrast, in domains like medicine characterized by limited data availability, generative models such as DGMs struggle to accurately represent the intricate relationships among features. Consequently, synthetic data generated under these circumstances often diverge significantly from the true data distribution. To address this challenge, [7] proposes incorporating an artificial inductive bias into DGM training.

Given a dataset comprising N samples, each characterized with C features, the methodology employs a DGM parameterized by θ to learn the distribution p_θ. The main objective of STDG involves generating synthetic samples x_g from p_θ that closely resemble the characteristics of real data x_r. While DGMs perform admirably in large sample datasets ($N \gg C$), their performance diminishes in data-limited scenarios, resulting in low synthetic data quality x_g. The methodology introduces an artificial inductive bias generator to enhance the quality of synthetic data when working with limited samples. This final STDG process unfolds through a dual-stage approach:

- *Inductive Bias Generation:* Initially, a DGM p_θ (potentially low-quality) is trained on the limited real data x_r to produce an initial set of synthetic data points x_g. These data points are then used to generate the inductive bias, resulting in an initial set of model weights θ_0.
- *Guided Training for Enhanced Synthetic Data:* Using θ_0 as the induced bias, a subsequent DGM $p_{\hat{\theta}}$ is trained. This guided training leverages θ_0 to generate enhanced synthetic data $\hat{x}_g$, which more closely resembles the characteristics of the real data x_r.

The methodology's essence lies in using data generated by an initial, potentially suboptimal DGM to form an initial weight set θ_0, facilitating improved convergence in the final model. Inductive biases are instrumental in guiding learning processes, particularly in scenarios with limited data. Effective bias incorporation leverages domain-specific insights, although tabular data often lack domain-specific knowledge.

1) Generating the Inductive Bias: Two main paradigms are explored to generate the θ_0 parameters:

- *Transfer learning [40]:* This approach enhances learning within the target domain (synthetic data) by transferring knowledge from a related context domain (real data). The methodology evaluates two distinct techniques:
  - *Pre-training:* Pre-training introduces an inductive bias into a model by leveraging a pre-trained model on a context domain to enhance performance on a target domain. This involves training an initial DGM $p_{\theta_{pt}}$ on pre-existing synthetic data x_g, generated from an initial DGM p_θ. This amount of synthetic data helps avoid overfitting. The optimal weights θ^*_{pt} from $p_{\theta_{pt}}$ are then used as initial weights θ_0 to train a new generative model $p_{\hat{\theta}}$. In other words, this approach, similar to data augmentation, uses synthetic data x_g to train a DGM $p_{\theta_{pt}}$. The goal is to use the information encoded in $p_{\theta_{pt}}$ to establish initial weights for the final DGM $p_{\hat{\theta}}$ that will be trained on real data x_r. Subfigure (a) from Fig. 1 illustrates this pre-training process, where the initial weights from synthetic data training enhance the model's performance on real data.
  - *Model-averaging:* This technique is a statistical approach that combines multiple models to enhance performance and better estimate uncertainty instead of selecting a single 'best' model. Model-Averaging maximizes information use and balances flexibility with overfitting. The methodology focuses particularly on models like VAEs sensitive to initial conditions. Multiple training runs (seeds) are typically conducted, and suboptimal seeds might be discarded. Instead, they use those varied seeds to create an artificial inductive bias by averaging their parameters. If S seeds are trained, resulting in parameters θ_S, the inductive bias θ_0 is calculated as

$$\theta_0 = \frac{1}{S}\sum_{s=1}^{S}\theta_s. \tag{1}$$

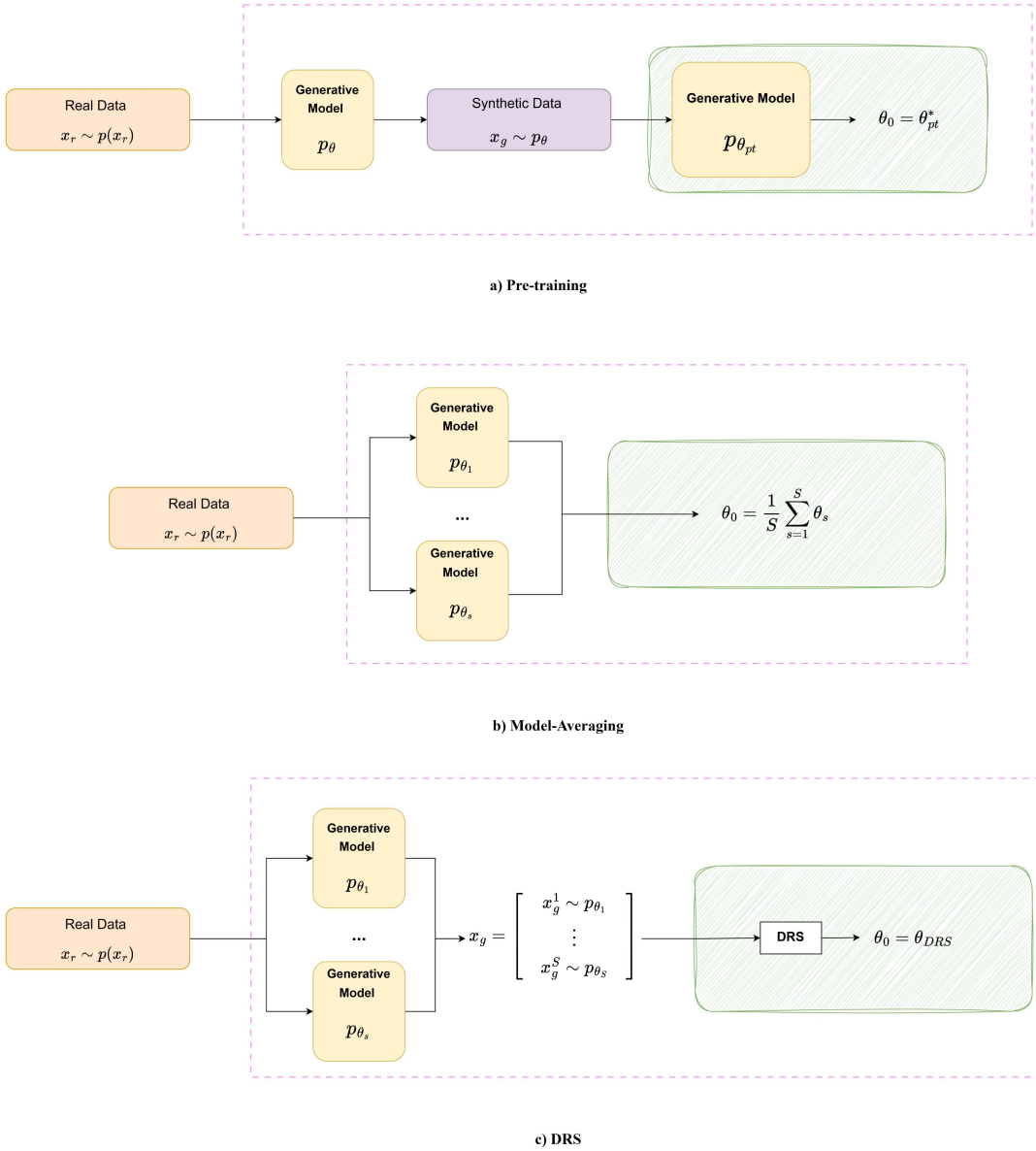


Fig. 1. Inductive bias generation techniques: (a) Pre-training on synthetic samples to initialize parameters; (b) Model-Averaging from multiple seeds to obtain initial parameters; (c) DRS combining synthetic data from multiple seeds for fine-tuning.

This method is computationally efficient, leveraging pre-computed weights to form a robust inductive bias, potentially improving model performance. Subfigure (b) from Fig. 1 illustrates this process, where the average weights from multiple seeds serve as the initial parameters for fine-tuning with real data.

Meta-learning [48]: This paradigm enables models to 'learn how to learn,' adeptly adapting to new tasks with minimal data. This approach leverages knowledge from multiple related tasks, allowing the models to dynamically adjust their learning strategies based on past experiences, thus requiring less data for optimal performance on similar tasks. In single-task ($\mathcal{T}$) learning, the goal is to find an optimal task-specific parameter θ^* that minimizes the loss function $\mathcal{L}$. The loss function depends on the chosen DGM, such as ELBO for VAEs. Meta-Learning extends this to a setting where a model is trained across a distribution of tasks $p(\mathcal{T})$, aiming to develop a general-purpose algorithm that performs well on new tasks with minimal data. In essence, meta-learning can be viewed as an extension of hyperparameter optimization. Here, the hyperparameter of interest, often called a meta-parameter, is shared across many tasks. This allows the model to capture the underlying structure of similar problems and apply that knowledge to new ones. The original methodology uses the multi-seed training configuration of certain DGMs, such as VAEs. Each S different seeds from VAE training is treated as a distinct task, forming a meta-learning framework that enhances the model's ability to generalize and adapt to new tasks efficiently. The paper explores techniques like

Model-Agnostic Meta-Learning (MAML) [20], which identifies an initialization parameter θ_{MAML} optimized across a set of training tasks. MAML optimizes θ_{MAML} such that it can be fine-tuned efficiently for new tasks by performing a few gradient updates on a small dataset. Formally, MAML solves the bi-level optimization problem:

$$\theta_{MAML} = \arg\min_{\theta} \mathop{\mathbb{E}}_{\mathcal{T}_b \sim p(\mathcal{T})} \mathcal{L}_b(\mathcal{D}_b^{(v)}; \theta_b^*), \tag{2}$$

subject to:

$$\theta_b^* = \theta - \alpha \nabla_\theta \mathcal{L}_b(\mathcal{D}_b^{(t)}; \theta), \tag{3}$$

where α is the learning rate for task-specific adaptation, $\mathcal{D}_b^{(t)}$ defines the training data used to adapt the model to task $\mathcal{T}_b$, and $\mathcal{D}b^{(v)}$ denotes the validation data used to evaluate the adapted model. The intuition behind MAML is to learn an initialization θ_{MAML} that quickly adapts to new tasks using $\mathcal{D}_b^{(t)}$, but this requires a diverse set of tasks and corresponding validation datasets $\mathcal{D}_b^{(v)}$ for adequate training. While effective, this process requires training across multiple diverse tasks, making it computationally expensive due to the need for repeated inner-loop gradient updates across a large task distribution.

Due to this limitation, we opted to use Domain Randomized Search (DRS) [22], a computationally efficient alternative that avoids the complexity of MAML's bi-level optimization. Unlike MAML, which optimizes for a meta-initialization θ_{MAML} by explicitly computing task-specific adaptations, DRS instead trains a model on the combined data from all tasks $\mathcal{T}_S \sim p(\mathcal{T})$, forming a single learning problem rather than a nested optimization. This eliminates the need for explicit inner-loop updates and significantly reduces computational costs. Formally, DRS solves:

$$\theta_{DRS} = \arg\min_{\theta} \mathop{\mathbb{E}}_{\mathcal{T}_S \sim p(\mathcal{T})} \mathcal{L}(\mathcal{D}; \theta), \tag{4}$$

where $\mathcal{D}$ represents the aggregated synthetic dataset from all training seeds S. This formulation allows DRS to approximate MAML by leveraging task diversity without explicitly performing meta-learning updates. Instead of learning an optimal initialization that requires task-specific fine-tuning, DRS finds a parameter θ_{DRS} that generalizes well across all tasks without requiring per-task adaptation.

Fig. 1(c) illustrates the DRS approach, where data from multiple seeds is used to train a shared model, initializing θ_{DRS} for fine-tuning on real data. Compared to MAML, this reduces computational overhead by replacing multiple task-specific optimizations with a single aggregated optimization, making it particularly suited for scenarios where the number of tasks is small [22].

Both MAML and DRS offer complementary trade-offs between optimization complexity and computational cost. While MAML achieves a theoretically more precise initialization, it requires many tasks to be effective, making it impractical for settings like ours, where each seed represents a task and S is relatively small (e.g., $S \approx 10$). DRS, by contrast, provides an efficient approximation to MAML that performs well with limited tasks while maintaining a significantly lower computational burden.

2) Deep Generative Model: This study utilizes a VAE-based model with a BGM [5], as proposed by [7], due to its proven effectiveness and flexibility. While the original methodology also included the CTGAN model, our focus on the VAE aligns with its demonstrated superiority and variability when trained with different seeds, a key aspect of generating inductive bias.

The VAE-based model combines the latent space learning of VAEs with the BGM's ability to model complex distributions. Unlike prior models like TVAE [55], which assume an isotropic Gaussian latent space, this approach recognizes that real-world data often exhibit more intricate dependencies. The VAE loss function incorporates the log-likelihood of reconstructed data and the Kullback-Leibler (KL) divergence, regularizing the latent space z. However, when the latent space deviates from the Gaussian assumption, sampling directly from it can produce suboptimal synthetic data. To address this, the BGM models z as a mixture of up to K Gaussian distributions, effectively capturing complex distributions and improving synthetic data quality. By leveraging the VAE's inherent variability, training with different seeds produces diverse outcomes, forming the foundation for the inductive bias in the generative methodology. This ensures the reliability and effectiveness of the STDG process, particularly for medical applications where accurate SDG is critical.

B. Validation Methodology

The approach proposed by [7] emphasizes divergences for validating synthetic data, building on [4]. Divergence calculations rely on a probabilistic discriminator network to estimate the density ratio between real and synthetic data distributions. This estimator captures discrepancies effectively, addressing the lack of standardized validation methods in STDG. Current similarity validation techniques often assess variables independently [25], overlooking inter-variable correlations and complex, non-linear relationships. To address this, the methodology employs two key metrics: KL and Jensen-Shannon (JS) divergences. KL divergence measures the difference between two probability distributions, quantifying the amount of information lost when one distribution is used to approximate the other. JS divergence, a symmetrized and bounded version of KL divergence, offers advantages in interpretability and robustness. Specifically, JS divergence provides a value between 0 and 1, making it easy to interpret, and it effectively considers complex correlations between variables, providing a more comprehensive similarity assessment.

However, in medical contexts, where practical application is critical, similarity validation alone is insufficient. Clinical utility validation is essential to assess whether synthetic data can perform specific tasks effectively. For this, we validate synthetic data for classification and SA tasks.

- *Classification tasks:* A Multi-Layer Perceptron (MLP) classifier is used for its flexibility and robustness in handling diverse classification problems. The accuracy score is the evaluation metric, measuring the proportion of correctly classified instances.
- *SA tasks:* The SAVAE model [6] is used for its superior performance over classic models like Cox Proportional Hazards (CoxPH) [18] and other DL models. Evaluation metrics include the concordance index (C-index) [3], which quantifies the concordance between predicted and observed survival times, and the Integrated Brier Score (IBS) [13], which assesses the precision of survival predictions over time. Additionally, Kaplan-Meier (KM) curves [29] visualize and compare survival probabilities, providing an intuitive method to evaluate similarities between real and synthetic data distributions.

Clinical utility validation is conducted across three cases to compare synthetic data performance:

1) *Real case:* Metrics obtained by training and validating on real data serve as the upper-bound benchmark.
2) *Synthetic case:* Metrics calculated by training on synthetic data and validating on real data.
3) *Synthetic fine-tuned case:* Metrics obtained by training synthetic data, fine-tuning on a separate real dataset, and validating the same real data used in the other cases.

This approach ensures consistent validation using the same real dataset, enabling direct comparison across cases and assessing the impact of synthetic data on utility tasks.

Clinical utility validation is crucial, but we can also have a better vision of how well the synthetic data are generated by changing the main task of the dataset. Therefore, to extend the scope of validation, additional tests are conducted by altering the main tasks of the datasets:

- *Classification datasets:* Target labels are modified to assess the synthetic data's adaptability to new classification tasks.
- *SA datasets:* Variables other than time, selected for their medical relevance, are used as targets to explore alternative predictive capabilities.

This comprehensive approach evaluates whether synthetic data can effectively support tasks beyond its original purpose (e.g., instead of predicting A, we want to predict B). It demonstrates the flexibility of synthetic datasets in addressing various clinical questions, enhancing their value and applicability in medical research. This method provides a broader clinical utility validation and helps accumulate evidence supporting the effectiveness of STDG.

The dual validation strategy (combining similarity and clinical utility) ensures that synthetic data approximates real data distributions and proves practical for specific medical applications. Similarity validation assures the fidelity of generated data, while clinical utility validation demonstrates task-specific effectiveness. This is crucial in medical settings where synthetic data may need to support diverse tasks beyond initial expectations. By incorporating both validation approaches, this methodology establishes synthetic data's reliability for real-world use and its potential to adapt to evolving research needs.

TABLE I
MEDICAL DATABASES USED IN EXPERIMENTS

Dataset	Number of samples	Number of features	Data types	Task
Heart	253,680	22	Binary, discrete	Classification
Metabric [43]	1,904	11	Binary, continuous	SA
Gbsb [21], [47]	2,232	9	Binary, continuous, discrete	SA
Nwtco [12]	4,028	8	Binary, discrete	SA

III. RESULTS

A. Medical Data

Medical data are highly heterogeneous, varying in sample size, features, and data types (e.g., binary, continuous, categorical). This study evaluates STDG methodologies across diverse applications, focusing primarily on cancer datasets for SA due to their complexity, variability, and clinical relevance. These datasets, ideal for time-to-event predictions, reflect real-world challenges of scarcity and heterogeneity. Classification datasets with abundant samples provide benchmarks for assessing performance under optimal and limited data conditions. As summarized in Table I, the selected databases span diverse types and dimensionalities, enabling a comprehensive evaluation of the methodology's robustness.

- *Classification dataset:* The Heart dataset[1] (253,680 samples) was selected for its substantial sample size [54]. This choice allows us to replicate experimental conditions from [7] and evaluate the methodology's performance under data-abundant conditions. This dataset establishes an upper-bound benchmark for comparison with SA datasets, typically containing fewer samples.
- *SA datasets:* Three SA datasets (Metabric, Gbsg, and Nwtco) were included to evaluate the methodology under real-world scarce data conditions. These datasets, downloaded from the Pycox package [31] and the SAVAE repository [6], are cancer-related and do not exceed 5,000 samples. Their smaller sample sizes allow us to test the methodology's robustness in preserving data similarity and generating reliable synthetic data under challenging conditions.

Our dual approach ensures a comprehensive assessment of the methodology's generalizability and effectiveness. Classification datasets provide insights into performance with abundant data, while SA datasets test adaptability to scarce and complex datasets. By addressing these distinct challenges, we aim to demonstrate the methodology's ability to generate synthetic data that maintain data similarity and practical utility across diverse medical contexts.

Additional experiments using other classification and SA datasets, detailed on our https://github.com/Patricia-A-Apellaniz/medical_low_sample_generator *repository*, further validate the methodology's applicability. Through this evaluation, our study highlights the potential of STDG techniques to address the complexities of medical data, contributing valuable insights for advancing SDG in healthcare.

[1]https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease Heart Dataset

B. Experimental Setting

Parameters were configured to balance complexity, computational efficiency, and data quality to implement the VAE-based model for STDG. The latent space dimensions were tailored to dataset types: 20 for classification datasets with more features and 10 for SA datasets with fewer features. Each hidden layer comprised 256 neurons, and the model was trained with ten different seeds, a batch size of 256, and up to 10,000 epochs, using early stopping to prevent overfitting. Other parameters for the VAE and the BGM models adhered to defaults outlined in the original methodology.

For clinical utility validation, the MLP classifier consisted of three dense layers (256, 64, and 32 neurons) with leaky ReLU activation, batch normalization, and dropout for regularization. The SAVAE model for SA was implemented with a latent dimension of 10 and hidden layers of 256 neurons, trained for 5,000 epochs using early stopping and a batch size of 256.

All models were cross-validated with five splits, ensuring robust and unbiased performance estimates. This is especially important for small datasets. This approach mitigates overfitting and allows comprehensive evaluation across data splits.

Three key experimental scenarios were designed to evaluate the methodology under varying data availability conditions:

1) *'Big data' scenario:* This optimistic scenario featured sufficient samples to train the DGM, with $N = 10,000$ for classification datasets and 80% of the total data for SA datasets. No inductive bias techniques were applied, establishing an upper-bound benchmark for divergences.
2) *'Low data' scenario:* This realistic scenario featured a limited sample size ($N = 100$) to assess baseline performance. The goal was to quantify potential gains from applying the proposed methodology under data-scarce conditions.
3) *'Pre-train,' 'AVG,' and 'DRS' Scenarios:* In the 'Low Data' scenario, advanced methods like Pre-training ('Pre-train'), Model-Averaging ('AVG') and DRS were applied to evaluate their effectiveness in generating high-quality synthetic data with limited samples.

To calculate divergences, 7,500 samples were used for training, and 1,000 for estimating the density ratio, consistent with [7]. While feasible for classification datasets with abundant data, this approach is less applicable to SA datasets, which typically have fewer samples. These experiments provided insights into the methodology's performance and divergence estimation under constrained data conditions.

Finally, p-values were used to rigorously assess differences in validation metrics, providing statistical insights into model performance across varying scenarios. Appendix B presents the p-values obtained for each dataset, comparing the 'Low-data' scenario to the different technique-enhanced scenarios. Given the large number of p-values generated, Holm's adjustment [27] was applied to control the family-wise error rate and ensure meaningful comparisons. The adjusted p-values are also provided in Appendix B, allowing for a clearer interpretation of the statistical significance of the results.

TABLE II
VALIDATION RESULTS FOR THE HEART DATASET UNDER DIFFERENT SCENARIOS: 'BIG DATA' ($N = 10,000$) REFLECTS OPTIMAL CONDITIONS, WHILE 'LOW DATA' ($N = 100$) REPRESENTS A MORE CHALLENGING SETTING FOR STDG

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	KL	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.096 (0.057)	0.171 (0.056)	0.631 (0.018)	0.615 (0.021)	0.629 (0.021)
Low data	0.852 (0.002)	6.642 (0.463)	0.601 (0.046)	0.640 (0.062)	0.636 (0.022)
Pre-train	**0.788 (0.004)**	**4.575 (0.242)**	N/A	0.687 (0.011)	0.692 (0.011)
AVG	**0.748 (0.008)**	**4.300 (0.135)**	N/A	0.696 (0.024)	0.665 (0.021)
DRS	**0.767 (0.017)**	**3.691 (0.122)**	N/A	0.649 (0.045)	0.661 (0.035)

Additional experiments, code, and data are openly available on our https://github.com/Patricia-A-Apellaniz/medical_low_sample_generator *repository* to promote replication and further research.

C. Experiments

Validation results are presented in tables and consist of two key components: similarity validation (using KL and JS divergences) and clinical utility validation (results across training and testing configurations). Lower values in the similarity validation indicate better similarity, reflecting overall dataset fidelity. Values in **bold** indicate significant improvements in divergences. Clinical utility configurations include training and validating with real data (Real *metric*), training with synthetic data and validating with real data (Synth. *metric*), and training with synthetic data, fine-tuning with real data, and validating with real data (Synth. Fine-Tuned *metric*). Higher accuracy or C-index and lower IBS values indicate better utility, focusing on preserving key task-relevant relationships. In this case, **bold** values indicate a significant advantage, while * indicates a disadvantage. All results are expressed as *mean (std)*.

1) Classification Datasets: The classification experiments focus on the Heart dataset. Table II summarizes its results. The 'Big data' scenario, with $N = 10,000$ samples, achieves the lowest JS divergence (0.096 ± 0.057), representing the optimal condition. In contrast, the 'Low data' scenario, with $N = 100$ samples, exhibits a significantly higher JS divergence (0.852 ± 0.002), reflecting the challenges of limited data. Advanced techniques, such as Model-Averaging and DRS, reduce divergences effectively, though they still fall short of the 'Big data' scenario, highlighting the persistent difficulty of STDG under data scarcity. A similar pattern is observed for KL divergences. This finding underscores that similarity validation results heavily depend on the number of samples used to generate synthetic data [4].

In clinical utility validation, three scenarios are analyzed: Real Acc. as the benchmark, Synth. Acc., and Synth. Fine-Tuned Acc. Results from the 'Big data' scenario serve as the upper bound, while the 'Low data' scenario highlights the methodology's challenges. In the benchmark scenario, accuracy metrics such as Real Acc. (0.631 ± 0.018) are slightly higher than in the 'Low data' scenario (0.601 ± 0.046). In this dataset, we observe in the benchmark scenario that when only a few real data samples are used, the accuracy obtained (0.601 ± 0.046) overlaps with

TABLE III
VALIDATION RESULTS FOR THE SA DATASETS UNDER DIFFERENT SCENARIOS: 'BIG DATA' ($N = 1,524$, $N = 1,786$, $N = 3,223$, 80% OF THE METABRIC, GBSG, NWTCO DATA, RESPECTIVELY) REFLECTS OPTIMAL CONDITIONS, WHILE 'LOW DATA' ($N = 100$) REPRESENTS A CHALLENGING SETTING FOR STDG

Dataset	Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Metabric	Big data	0.633 (0.035)	0.622 (0.035)	0.626 (0.035)	0.183 (0.028)	0.196 (0.028)	0.185 (0.028)
	Low data	0.595 (0.037)	0.589 (0.040)	0.587 (0.041)	0.194 (0.029)	0.199 (0.029)	0.200 (0.029)
	Pre-train	N/A	0.614 (0.037)	0.617 (0.038)	N/A	0.184 (0.028)	0.186 (0.028)
	AVG	N/A	0.613 (0.035)	0.615 (0.036)	N/A	0.182 (0.028)	0.183 (0.028)
	DRS	N/A	0.614 (0.036)	0.604 (0.040)	N/A	0.183 (0.028)	0.184 (0.028)
Gbsg	Big data	0.683 (0.032)	0.690 (0.031)	0.685 (0.031)	0.193 (0.026)	0.184 (0.026)	0.192 (0.026)
	Low data	0.638 (0.041)	0.598 (0.053)	0.595 (0.052)	0.208 (0.028)	0.232 (0.034)	0.228 (0.033)
	Pre-train	N/A	0.656 (0.033)	0.657 (0.033)	N/A	0.213 (0.028)	0.211 (0.029)
	AVG	N/A	0.647 (0.043)	0.641 (0.051)	N/A	0.214 (0.028)	0.217 (0.029)
	DRS	N/A	0.654 (0.036)	0.672 (0.032)	N/A	0.198 (0.027)	0.196 (0.027)
Nwtco	Big data	0.694 (0.023)	0.682 (0.028)	0.683 (0.025)	0.112 (0.016)	0.111 (0.016)	0.111 (0.015)
	Low data	0.587 (0.044)	0.549 (0.028)	0.542 (0.028)	0.139 (0.023)	0.141 (0.017)	0.136 (0.017)
	Pre-train	N/A	0.590 (0.025)	0.596 (0.025)	N/A	0.152 (0.019)	0.146 (0.018)
	AVG	N/A	0.591 (0.026)	0.593 (0.028)	N/A	0.138 (0.019)	0.142 (0.017)
	DRS	N/A	**0.611 (0.033)**	**0.594 (0.037)**	N/A	0.133 (0.018)	0.131 (0.018)

the 'Big data' scenario (0.631 ± 0.018). This pattern persists in the other two cases where the methodology is not used. The transfer-learning techniques do not show significant advantages or disadvantages in the Synth. and Synth. Fine-tuned cases compared to the benchmark accuracies obtained in the Real Acc. case.

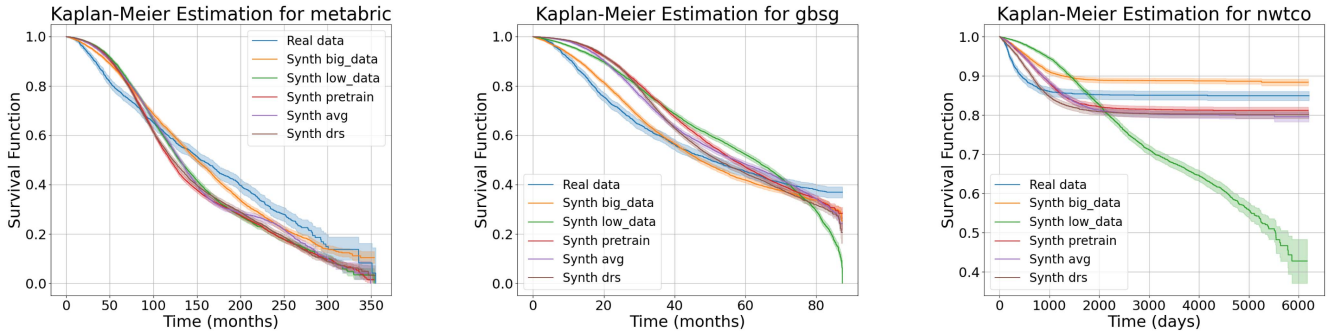
2) SA Results: Classification data demonstrated that the methodology improves SDG, particularly regarding divergences. Additionally, we reiterated that divergences are a robust and reliable metric to validate SDG as proved in [4]. Building on this, SA experiments focus on datasets with limited sample sizes to test the robustness of the methodology in real-world, data-scarce conditions. These cancer-related datasets, chosen for their scarcity and heterogeneity, present a challenging but realistic testbed for evaluating synthetic data quality. Due to the limited sample sizes in SA datasets, accurate calculation of divergences is not feasible, as inadequate data affects the training of DGMs. Thus, similarity validation is omitted for SA datasets, and the focus shifts exclusively to clinical utility validation. Clinical utility metrics, including C-index and IBS, assess the methodology's practical applicability under previously defined scenarios.

Table III presents the C-index and IBS for the three cases across each scenario. The results highlight two key observations: (1) there is no significant loss in any metric when comparing the 'Big data' scenario to the 'Low data' scenario, aligning with the classification data results in clinical utility validation, and (2) there is no difference in performance metrics when utilizing the methodologies, consistent with previous experiments. This confirms that clinical utility validation alone is insufficient to evaluate the quality of synthetic data. We hypothesize that using the methodology would yield better divergence metrics, in line with our previous experiment and the findings in [7]; however, note that we cannot reliably assess divergences with a low number of samples.

KM estimations were performed to supplement clinical utility validation. With Confidence Intervals (CIs), these survival curves visually compare survival probabilities for real and synthetic data across scenarios, offering additional insights into synthetic data quality. Fig. 2 highlights notable trends, particularly in the Nwtco dataset. KM curves generated with 'Big data' closely align with those from real data, serving as the benchmark. In the 'Low data' scenario, curves deviate significantly from the benchmark without the methodology, with broader CIs. Methodology-enhanced scenarios, such as DRS, produce KM curves that converge toward the upper bound, narrowing the gap between synthetic and real data. The Gbsg dataset exhibits similar trends, particularly in later survival time regions, where methodology-applied curves diverge less from the benchmark than the 'Low data' curves. These results confirm that the methodology effectively improves synthetic data quality under limited data conditions.

3) Different Data Utility Results: This section evaluates the robustness of synthetic data generated using STDG by modifying original tasks for classification and SA datasets. Target labels are altered in classification datasets, while categorical variables replace survival time in SA datasets, transforming them into classification problems. This approach assesses the adaptability and reliability of synthetic data across diverse clinical applications. One of STDG's primary goals is to create datasets that can be repurposed for tasks beyond their original intent. Traditional clinical utility validation focuses on relationships between covariates and target variables. However, testing whether synthetic data retains utility when target variables change is critical, as it mirrors real-world scenarios where datasets are used for different studies.

Tables IV and V summarize the results for each dataset. Accuracy values were compared across two scenarios: (1) training with $N = 100$ samples of synthetic data generated without the methodology ('Low data') and (2) training with $N = 100$ samples of synthetic data generated using the DRS technique and validating with real data ('DRS'). The first subtable shows the Heart dataset results, where different features were used as target labels. These features were binary or categorical, with



(a) Metabric dataset (b) Gbsg dataset (c) Nwtco dataset

Fig. 2. KM estimations with CIs for real and synthetic data across scenarios. Upper bounds (blue/orange) represent survival probabilities for real and synthetic data from many samples. Synthetic data from the proposed methodology (red/purple/brown) converge towards upper bounds, while the lower bound (green) deviates significantly.

TABLE IV
CLINICAL UTILITY VALIDATION RESULTS FOR THE HEART DATASET, COMPARING ACCURACY BETWEEN THE 'LOW DATA' CASE ($N = 100$ SAMPLES) AND THE DRS METHODOLOGY APPLIED TO THE LOWER BOUND CASE

Feature	Low data	DRS	*Feature*	Low data	DRS
HighBP	0.664 (0.005)	**0.695 (0.005)**	*Veggies*	0.519 (0.167)	0.609 (0.069)
HighChol	0.603 (0.008)	0.609 (0.004)	*HvyAlcoholConsump*	0.456 (0.294)	0.934 (0.038)
CholCheck	0.665 (0.146)	0.493 (0.199)	*AnyHealthcare*	0.629 (0.048)	**0.755 (0.012)**
Smoker	0.599 (0.006)	0.604 (0.008)	*NoDocbcCost*	0.728 (0.047)	0.639 (0.011)
Stroke	0.768 (0.280)	0.664 (0.010)	*DiffWalk*	0.799 (0.019)	0.795 (0.003)
PhysActivity	0.671 (0.017)	0.675 (0.004)	*Sex*	0.539 (0.016)	0.528 (0.018)
Fruits	0.588 (0.022)	0.585 (0.028)	*HeartDiseaseorAttack*	0.679 (0.123)	0.671 (0.055)

fewer than ten classes. The results show that the DRS technique improves accuracy compared to the 'Low data' scenario. Notable improvements include using *HighBP* and *AnyHealthcare* as target labels, where accuracy increased significantly. For example, *AnyHealthcare* improved from 0.629 to 0.755 with the DRS technique.

IV. DISCUSSION

This study addresses the critical limitations of current STDG techniques, particularly their reliance on large datasets, which is impractical for many medical applications. This methodology introduces an artificial inductive bias through transfer learning and meta-learning, enabling the generation of high-quality synthetic data in data-scarce scenarios. Existing STDG models, such as CTGAN and MedGAN, depend on datasets with tens of thousands of samples to generate realistic synthetic data [17], [55]. While these approaches achieve high fidelity in large-scale datasets, they are ill-suited for medical contexts characterized by data scarcity, such as rare diseases or small-scale clinical trials. The application of this methodology [7] bridges the gap by extending STDG capabilities to small, heterogeneous datasets, balancing fidelity with task-specific clinical utility. Unlike traditional models, which often prioritize overall distributional fidelity, our approach emphasizes the accurate generation of task-critical variables, ensuring clinical relevance.

In classification and SA datasets, the methodology yields notable improvements in divergences under low-data scenarios, indicating better alignment between synthetic and real data distributions. However, clinical utility validation without the methodology often achieves results comparable to the benchmark scenarios, suggesting that the accurate generation of critical variables for classification can compensate for discrepancies in the overall distribution. This behavior might be explained by the accurate generation of key variables, which contain sufficient information for correct classification even if other variables are not well-represented. Consequently, despite high divergence values indicating discrepancies in the synthetic joint distribution, utility metrics for synthetic data can still closely approximate those achieved with real data. In Appendix A, we provide further insights into this behavior, discussing how the generation of these critical variables can suffice for clinical utility validation, even when the joint distribution of the synthetic data is suboptimal. This underscores the importance of combining similarity and utility validation to assess synthetic data quality comprehensively.

Furthermore, the findings suggest that the methodology effectively models complex relationships between features, enhancing data utility for classification tasks. Even in scenarios where no significant improvement is observed, the methodology does not negatively affect performance. In SA datasets, categorical variables were used as target labels instead of survival time to assess the data's utility in classification tasks. The results align with those from the classification datasets, with the DRS method consistently improving accuracy for specific features. For example, in the Nwtco dataset, the feature *instit_2* increased accuracy from 0.605 ± 0.034 to 0.726 ± 0.046. Similarly, the Gbsg dataset's feature *x2* improved from 0.873 ± 0.007 to 0.890 ± 0.004. While some features, such as *x7* in the Metabric dataset, experienced slight decreases, these cases are rare, and the overall trend indicates improved or maintained performance with the methodology. By validating the application of the methodology through these transformed tasks, we demonstrate that using the generation methodology either yields benefits in specific scenarios or does not negatively affect the outcomes. This evidence supports that incorporating an inductive bias for STDG in low-sample datasets enhances the modeling of

TABLE V
CLINICAL UTILITY VALIDATION RESULTS FOR SA DATASETS, COMPARING ACCURACY BETWEEN THE 'LOW DATA' CASE ($N = 100$ SAMPLES) AND THE DRS METHODOLOGY APPLIED TO THE LOWER BOUND CASE

Metabric			Gbsg			Nwtco		
Feature	**Low data**	**DRS**	*Feature*	**Low data**	**DRS**	*Feature*	**Low data**	**DRS**
x4	0.676 (0.012)	0.652 (0.018)	x0	0.595 (0.012)	0.607 (0.021)	in.subcohort	0.678 (0.150)	0.578 (0.216)
x5	0.578 (0.015)	0.579 (0.020)	x1	0.326 (0.071)	0.322 (0.064)	instit_2	0.605 (0.034)	**0.726 (0.046)**
x6	0.748 (0.003)	0.740 (0.023)	x2	0.873 (0.007)	**0.890 (0.004)**	histol_2	0.714 (0.153)	0.695 (0.037)
x7	0.795 (0.013)	0.724 (0.017)*	event	0.805 (0.006)	0.809 (0.001)	study_4	0.782 (0.008)	0.763 (0.012)
event	0.696 (0.004)	0.701 (0.006)				event	0.687 (0.029)	**0.788 (0.005)**

Higher values indicate better performance.

complex relationships between variables. Consequently, this approach improves the generation process, providing better utility and robustness for diverse clinical applications.

V. CONCLUSION

In DL, large amounts of data are ideal, but data are often scarce and heterogeneous in the medical field. This research applied the methodology proposed by [7] to STDG in medical environments, focusing on cancer-related SA datasets with limited samples. As a preliminary step, classification datasets with more samples were used to evaluate the methodology's performance by comparing results from subsets of data to those obtained using the entire dataset. The validation process combined similarity validation using divergences, particularly the interpretable and bounded JS divergence, and clinical utility validation. These assessments demonstrated that transfer learning and meta-learning techniques in STDG environments enhance model performance, especially under data-scarce conditions. Initially, the methodology was tested on classification datasets, where its application improved the generation process. Divergences proved to be robust metrics for comparing real and synthetic datasets, while clinical utility validation showed limited sensitivity, as metrics like accuracy remained stable across scenarios. Clinical utility validation metrics (C-index and IBS) showed minimal variation compared to upper-bound results for cancer-related SA datasets characterized by few samples. These findings confirmed that the synthetic data were sufficient SA tasks, though divergences were more reliable than utility metrics for validating STDG. When datasets were repurposed for alternative tasks, the methodology occasionally generated better synthetic data, supporting its ability to model complex relationships and produce data usable across different clinical applications. This research demonstrated that the methodology significantly improves STDG for medical applications with limited data, often making validation metrics statistically comparable to those in the 'Big data' scenario and outperforming the naive 'Low data' approach. JS divergence emerged as a reliable tool for comparing real and synthetic datasets, while clinical utility validation requires further refinement for low-sample scenarios. Consequently, this work provides the first empirical validation of the inductive bias methodology from [7] in the medical domain. By comparing against low-data baselines using standard DGMs, we show that our adapted methodology systematically improves synthetic data quality in challenging real-world settings. Despite its limitations, the methodology enables SDG for broader medical analyses.

Future work should focus on developing more robust validation metrics by combining alternative divergence measures with clinical utility assessments for a comprehensive evaluation framework. Collaboration with clinical experts is essential to refine the STDG process and ensure that the generated data meet medical standards. Establishing public repositories of high-quality synthetic datasets would facilitate further research while maintaining patient privacy and data security, promoting innovation in medical research. Additionally, future research should explore the validation of synthetic data using multi-modal datasets, which often integrate structured tabular data (EHRs, laboratory results), imaging (radiology, pathology), temporal data (disease progression, monitoring), and biological data (genomics, transcriptomics, proteomics, metabolomics, microbiome profiles). Evaluating the STDG approach across these modalities would help assess its ability to capture cross-modal dependencies and generate coherent synthetic representations. This extension is particularly relevant for precision medicine and clinical decision support systems, where multi-modal data fusion enhances diagnosis, treatment planning, and outcome prediction. Furthermore, a rigorous analysis of the relationship between sample size, dataset complexity, and validation metrics is needed. The minimum sample size for effective SDG depends on feature count, distributions, correlations, and overall dataset complexity. As validation metrics vary in sensitivity to sample size, future studies should establish quantitative guidelines for evaluating STDG in data-scarce scenarios. Future studies should systematically explore these factors to provide more precise recommendations on sample size requirements for effective STDG.

APPENDIX A
P-VALUES HOLM ADJUSTMENT

To ensure statistical rigor, we tested the null hypothesis (H_0): Performance metrics of the 'Low data' scenario are better than those where the methodology is applied. A p-value threshold of 0.01 determined significance, rejecting H_0 when below this threshold, indicating a significant improvement.

To address the increased Family-Wise Error Rate (FWER) due to multiple tests [49], [51], we applied the Holm-Bonferroni method, which adjusts p-values to control overall significance while offering more power than the traditional Bonferroni

TABLE VI
VALIDATION p-VALUE RESULTS FORMATTED AS 'ORIGINAL p-VALUE/ADJUSTED p-VALUE' AFTER HOLM ADJUSTMENT FOR MULTIPLE TESTING

Dataset	Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION	
		JS	KL	Synth Acc	Synth Fine-Tuned Acc
Heart	Pre-train	0.000 / **0.000**	0.000 / **0.000**	0.012 / 0.024	0.010 / 0.030
	AVG	0.000 / **0.000**	0.000 / **0.000**	0.006 / 0.019	0.032 / 0.064
	DRS	0.000 / **0.000**	0.000 / **0.000**	0.130 / 0.130	0.050 / 0.064

(a) Classification dataset

Dataset	Scenario	Synth CI	Synth Fine-Tuned CI	Synth IBS	Synth Fine-Tuned IBS
Metabric	Pre-train	0.431 / 1.000	0.380 / 1.000	0.603 / 1.000	0.643 / 1.000
	AVG	0.450 / 1.000	0.400 / 1.000	0.519 / 1.000	0.537 / 1.000
	DRS	0.433 / 1.000	0.710 / 1.000	0.542 / 1.000	0.578 / 1.000
Gbsg	Pre-train	0.468 / 1.000	0.430 / 0.859	0.814 / 1.000	0.886 / 1.000
	AVG	0.747 / 1.000	0.926 / 0.926	0.773 / 1.000	0.650 / 1.000
	DRS	0.525 / 1.000	0.184 / 0.551	0.572 / 1.000	0.479 / 1.000
Nwtco	Pre-train	0.894 / 1.000	0.712 / 1.000	0.358 / 1.000	0.635 / 1.000
	AVG	0.867 / 1.000	0.813 / 1.000	0.946 / 1.000	0.850 / 1.000
	DRS	0.355 / 1.000	0.781 / 1.000	0.636 / 1.000	0.526 / 1.000

(b) Survival analysis datasets

Bold values indicate statistically significant differences from the 'low data' scenario (null hypothesis rejected).

correction [10]. The adjustment, implemented using Python's *statsmodels* package, ensures consistent thresholds across tests, minimizing false positives. Original and adjusted p-values are summarized in Table VI.

Our analysis demonstrates that after applying the Holm adjustment, significant improvements were observed in the classification dataset, particularly in the JS similarity metric, where adjusted p-values consistently fell below 0.01. This confirms that the proposed methodology leads to statistically meaningful gains over the 'Low data' scenario in terms of data resemblance. In contrast, for the SA datasets, no significant improvements were observed in any metric. This result aligns with previous observations discussed in the main text.

APPENDIX B
DIVERGENCE BETWEEN SIMILARITY VALIDATION AND CLINICAL UTILITY VALIDATION

This section explores the observed divergence between similarity and clinical utility validation results.

A. Proposed Analysis

We analyzed classification datasets to examine how feature importance impacts accuracy using the MLP classifier. Since MLP lacks inherent feature importance measures, we employed SHAP (SHapley Additive exPlanations) to estimate each feature's contribution. The process involved:

1) Classification with the complete dataset with all available features.
2) Importance analysis using SHAP to rank the features.
3) Remove the feature with the lowest absolute SHAP value.
4) Reclassification using the reduced dataset.
5) Repeat steps 2–4 iteratively.

This experiment shows that classification accuracy remains stable even when less important features are removed, highlighting that only a subset is critical for high performance.

This explains the divergence between validation methods:

TABLE VII
CLASSIFICATION ACCURACY RESULTS FOR MLP MODELS TRAINED ON THE HEART DATASET WITH PROGRESSIVELY FEWER FEATURES

Feature removed	Number of features	Acc	Feature removed	Number of features	Acc
None	22	0.643	Stroke	11	0.640
AnyHealthcare	21	0.637	PhysHlth	10	0.608
MentHlth	20	0.630	Diabetes	9	0.619
Education	19	0.616	DiffWalk	8	0.612
CholCheck	18	0.632	Smoker	7	0.593
Veggies	17	0.638	Income	6	0.640
HvyAlcoholConsump	16	0.634	Sex	5	0.630
NoDocbcCost	15	0.645	HighBP	4	0.635
BMI	14	0.650	HighChol	3	0.650
Fruits	13	0.645	GentHlth	2	0.560
PhysActivity	12	0.655			

Accuracy (Acc) is reported for each configuration.

- *Similarity validation:* Assesses overall dataset quality, treating all features equally.
- *Clinical utility validation:* Focuses on task-critical features, performing well if these are accurately generated, regardless of other features' quality.

Thus, synthetic data may score poorly on similarity validation but still excel in clinical utility validation if key features are well-represented. This emphasizes the need for task-specific evaluation alongside global data quality assessments.

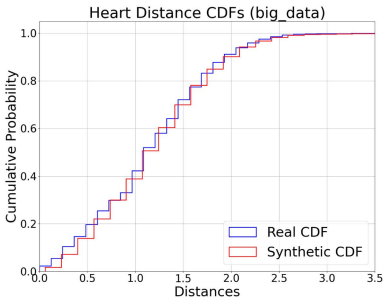
B. Results

This analysis uses the Heart dataset, a binary classification problem, to explore the relationship between feature importance and classification accuracy. Its simplicity enables a clear interpretation of SHAP values, supporting the proposed methodology. Table VII shows that MLP classification accuracy remains stable (around 0.650) as less important features are removed, with a noticeable drop only when fewer than five remain. This indicates that only a subset of features is crucial for performance, with most features contributing minimally to accuracy. These findings explain the divergence between validation methods. Similarity validation assesses the generation of all features, making it sensitive to global discrepancies. Conversely, clinical utility validation focuses on key features critical to the task, achieving high performance even if other features are poorly generated. This highlights the importance of combining both methods to comprehensively evaluate synthetic data, balancing dataset fidelity with task-specific performance.

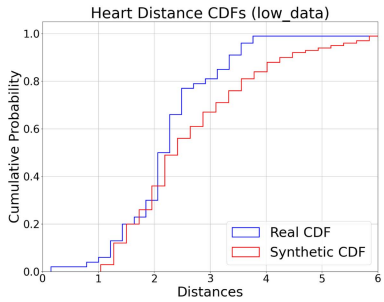
APPENDIX C
OVERFITTING PREVENTION AND EMPIRICAL VALIDATION

Our proposed generative model, based on VAEs, inherently mitigates overfitting through its regularized training objective. Specifically, including the KL divergence term in the ELBO encourages the approximate posterior distribution to remain close to the prior. This regularization prevents the latent space from collapsing into deterministic point estimates, promoting stochastic latent representations that reduce the likelihood of the model memorizing training data.

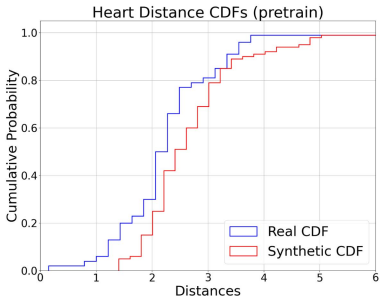
To empirically validate that our model does not overfit the training data, we conducted privacy assessments focusing on the similarity between real and synthetic samples. We calculated the minimum pairwise Euclidean distances between real-real



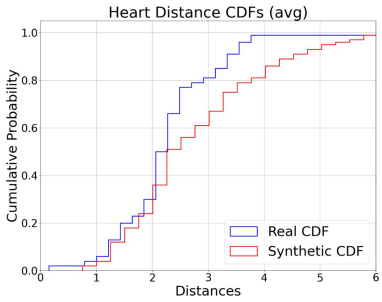
(a) Big data scenario. Significant difference observed (Wilcoxon $p - value = 1.30^{-100}$; KS $p - value = 2.26^{-11}$).



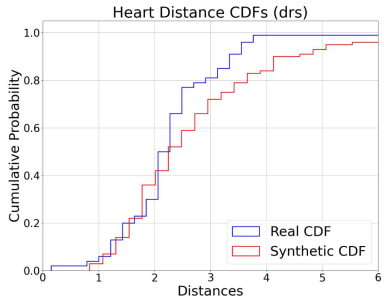
(b) Low data scenario. Significant difference observed (Wilcoxon $p - value = 2.33^{-9}$; KS $p - value = 0.0120$).



(c) Pre-train scenario. Significant difference observed (Wilcoxon $p - value = 1.52^{-8}$; KS $p - value = 0.0006$).



(d) AVG scenario. Significant difference observed (Wilcoxon $p - value = 2.38^{-10}$; KS $p - value = 0.0031$).



(e) DRS scenario. Significant difference observed (Wilcoxon $p - value = 1.94^{-7}$; KS $p - value = 0.0120$.

Fig. 3. Comparison of minimum pairwise distances between real-real samples and synthetic-real samples. The CDF plots show similar distributions, ensuring statistical resemblance while maintaining privacy.

and synthetic-real data pairs. The results, visualized through CDFs, indicate that the minimum distances between synthetic and real samples are consistently greater than those among real samples and are strictly positive. This suggests that the synthetic data do not replicate real data points. Furthermore, one-sided Wilcoxon and Kolmogorov-Smirnov (KS) tests confirmed that these differences are statistically significant (p-values < 0.05), reinforcing the conclusion that our generative process avoids overfitting.

These evaluations were performed on the Heart dataset for each scenario (Fig. 3), providing a comprehensive assessment of privacy preservation across different data distributions and experimental settings. The findings support the capability of our model to generate high-quality synthetic data without compromising privacy or exhibiting overfitting.

APPENDIX D
COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART GENERATORS

We provide a comparative analysis between the VAE-based model with a BGM (VAE-BGM) [5] and two widely recognized synthetic data generators: CTGAN and TVAE [55]. This evaluation focuses on both 'Big data' and 'Low data' scenarios to assess the models' performance across varying data availability conditions.

TABLE VIII

Scenario	Model	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
		JS	KL	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	VAE-BGM	0.096 (0.057)	0.171 (0.056)	0.631 (0.018)	0.615 (0.021)	0.629 (0.021)
	CTGAN	0.149 (0.001)	0.380 (0.042)		0.701 (0.007)	0.685 (0.016)
	TVAE	0.844 (0.001)	8.129 (0.140)		0.544 (0.035)	0.640 (0.009)
Low data	VAE-BGM	0.852 (0.002)	6.642 (0.463)	0.601 (0.046)	0.640 (0.062)	0.636 (0.022)
	CTGAN	0.705 (0.008)	2.787 (0.075)		0.681 (0.033)	0.696 (0.030)
	TVAE	0.933 (0.002)	10.430 (0.845)		0.605 (0.294)	0.667 (0.067)

In the classification task using the Heart dataset (Table VIII), the VAE-BGM model demonstrated superior performance in similarity validation under the 'Big data' scenario, achieving lower JS and KL divergences compared to CTGAN and TVAE. Although CTGAN exhibited lower divergence values in the 'Low data' scenario, these results may not be reliable due to the limited sample size, which increases the risk of overfitting and reduces the generalizability of the synthetic data. TVAE consistently showed higher divergence values, likely due to its assumption of a single Gaussian distribution in the latent space, which may not adequately capture the complexity of real-world data distributions. Regarding clinical utility validation in the Heart dataset, CTGAN achieved higher accuracy metrics in both 'Big data' and 'Low data' scenarios. However, the improved performance in the 'Low data' scenario should be interpreted with caution, as it may result from overfitting to the limited training

TABLE IX

Dataset	Scenario	Model	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBs	Synth Fine-Tuned IBS
Metabric	Big data	VAE-BGM	0.633 (0.035)	0.622 (0.035)	0.626 (0.035)	0.183 (0.028)	0.196 (0.028)	0.185 (0.028)
		CTGAN		0.584 (0.036)	0.589 (0.035)		0.223 (0.030)	0.214 (0.030)
		TVAE		0.611 (0.035)	0.611 (0.035)		0.258 (0.031)	0.222 (0.032)
	Low data	VAE-BGM	0.595 (0.037)	0.589 (0.040)	0.587 (0.041)	0.194 (0.029)	0.199 (0.029)	0.200 (0.029)
		CTGAN		0.597 (0.035)	0.596 (0.037)		0.191 (0.028)	0.192 (0.028)
		TVAE		0.532 (0.035)	0.527 (0.038)		0.259 (0.031)	0.244 (0.032)
Gbsg	Big data	VAE-BGM	0.683 (0.032)	0.690 (0.031)	0.685 (0.031)	0.193 (0.026)	0.184 (0.026)	0.192 (0.026)
		CTGAN		0.670 (0.031)	0.673 (0.031)		0.182 (0.025)	0.186 (0.026)
		TVAE		0.515 (0.033)	0.531 (0.033)		0.238 (0.028)	0.227 (0.028)
	Low data	VAE-BGM	0.638 (0.041)	0.598 (0.053)	0.595 (0.052)	0.208 (0.028)	0.232 (0.034)	0.228 (0.033)
		CTGAN		0.611 (0.037)	0.609 (0.033)		0.212 (0.027)	0.216 (0.029)
		TVAE		0.553 (0.034)	0.576 (0.044)		0.240 (0.028)	0.227 (0.028)
Nwtco	Big data	VAE-BGM	0.694 (0.023)	0.682 (0.028)	0.683 (0.025)	0.112 (0.016)	0.111 (0.016)	0.111 (0.015)
		CTGAN		0.682 (0.024)	0.673 (0.024)		0.147 (0.017)	0.133 (0.017)
		TVAE		0.640 (0.034)	0.654 (0.025)		0.175 (0.019)	0.138 (0.017)
	Low data	VAE-BGM	0.587 (0.044)	0.549 (0.028)	0.542 (0.028)	0.139 (0.023)	0.141 (0.017)	0.136 (0.017)
		CTGAN		0.674 (0.025)	0.665 (0.025)		0.136 (0.017)	0.139 (0.019)
		TVAE		0.636 (0.032)	0.607 (0.034)		0.264 (0.022)	0.234 (0.021)

data, potentially compromising the model's ability to generalize to unseen data. For the SA tasks involving the Metabric, Gbsg, and Nwtco datasets (Table IX), the VAE-BGM model consistently achieved lower IBS across most scenarios, indicating better calibration of survival probabilities. In terms of C-index, VAE-BGM generally outperformed CTGAN in the 'Big data' scenarios, suggesting more stable and reliable survival predictions. CTGAN's performance varied across datasets, showing superior and inferior results compared to VAE-BGM, which may reflect its sensitivity to specific data characteristics. TVAE consistently underperformed in IBS and C-index, reinforcing the limitations of its latent space assumptions in capturing complex SA patterns.

Overall, the VAE-BGM model offers a robust and reliable approach for STDG across various medical datasets and scenarios, balancing similarity to real data and clinical utility, particularly in settings with limited data. Nevertheless, to enhance the analysis, it would be valuable to explore the application of these techniques to DGMs not based on VAEs, such as CTGAN. However, this would necessitate adapting several processes due to architectural differences.

APPENDIX E
COMPUTATIONAL RESOURCES AND TRAINING TIME ANALYSIS

Experiments were conducted on a prosumer-grade workstation with an AMD Ryzen Threadripper PRO 5975WX processor, which features 32 cores and 64 threads. This processor operates at a base frequency of 3.6 GHz and can boost up to 4.5 GHz, providing substantial computational power for parallel processing tasks. The system was complemented with 128 GB of DDR4 RAM, ensuring ample memory for handling large datasets and complex computations. Storage configurations included a 1 TB NVMe SSD for fast data access and two 4 TB HDDs for additional storage needs. Although the workstation had an NVIDIA RTX 4090 graphics card, known for its high performance in graphics-intensive tasks, no GPU acceleration was employed during the experiments; all training and evaluation processes were executed solely on the CPU.

TABLE X
AVERAGE EXECUTION TIMES (IN SECONDS) ACROSS 10 RUNS WITH DIFFERENT SEEDS

Dataset	Big data	Low data	Pre-train	AVG	DRS
Heart	327.037	15.085	284.056	29.709	100.878
Metabric	47.805	11.398	120.753	36.126	349.818
Gbsg	53.749	16.380	114.122	42.594	329.455
Nwtco	58.155	9.310	115.064	29.760	387.097
Average	121.687	13.043	158.499	34.547	291.812

Despite the expressive nature of the proposed model, training times remained within practical limits. As summarized in Table X, the average execution times across 10 runs with different random seeds for various scenarios and datasets indicate that all training processes were completed in under 10 minutes. These reduced and manageable durations demonstrate that our method is computationally feasible even on standard hardware configurations.

ACKNOWLEDGMENT

The content reflects the authors' views only, and the EU or European Commission is not responsible for any use that may be made of it.

REFERENCES

[1] I. Al-Hurani, A. Alkhateeb, and S. Ikki, "An autoencoder and generative adversarial networks approach for multi-omics data imbalanced class handling and classification," 2024, *arXiv:2405.09756*.
[2] F. Alghanim et al., "Machine learning model for multiomics biomarkers identification for menopause status in breast cancer," *Algorithms*, vol. 17, no. 1, 2023, Art. no. 13.
[3] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statist. Med.*, vol. 24, pp. 3927–3944, 2005.
[4] P. A. Apellániz, A. Jiménez, B. A. Galende, J. Parras, and S. Zazo, "Synthetic tabular data validation: A divergence-based approach," *IEEE Access*, vol. 12, pp. 103895–103907, 2024.
[5] P. A. Apellaniz, J. Parras, and S. Zazo, "An improved tabular data generator with VAE-GMM integration," in *Proc. 32nd Eur. Signal Process. Conf.*, Lyon, France, 2024, pp. 1886–1890. doi: 10.23919/EUSIPCO63174.2024.10715230.

[6] P. A. Apellániz, J. Parras, and S. Zazo, "Leveraging the variational Bayes autoencoder for survival analysis," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 24567.

[7] P. A. Apellániz, A. Jiménez, B. A. Galende, J. Parras, and S. Zazo, "Artificial inductive bias for synthetic tabular data generation in data-scarce scenarios," *Neurocomputing*, 2025, Art. no. 131122.

[8] S. M. Bellovin, P. K. Dutta, and N. Reitinger, "Privacy and synthetic datasets," *Stanford Technol. Law Rev.*, vol. 22, no. 1, 2019. [Online]. Available: https://law.stanford.edu/wpcontent/uploads/2019/01/Bellovin_20190129.pdf

[9] A. Bohr, T. Altstidl, B. Eskofier, and E. Salin, "Perspective: Leveraging domain knowledge for tabular machine learning in the medical domain," in *Proc. 4th Table Representation Learn. Workshop*, 2025, pp. 143–155.

[10] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze*, vol. 8, pp. 3–62, 1936.

[11] A. Brauneck et al., "Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: Scoping review," *J. Med. Internet Res.*, vol. 25, 2023, Art. no. e41588.

[12] N. E. Breslow and N. Chatterje, "Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis," *J. Roy. Stat. Soc., Ser. C (Appl. Statist.)*, vol. 48, no. 4, pp. 457–468, 1999.

[13] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.

[14] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[16] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 493–497, 2021.

[17] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Proc. 2nd Mach. Learn. Healthcare Conf.*, 2017, pp. 286–305.

[18] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972.

[19] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[21] J. A Foekens et al., "The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients," *Cancer Res.*, vol. 60, no. 3, pp. 636–643, 2000.

[22] K. Gao and O. Sener, "Modeling and optimization trade-off in meta-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11154–11165.

[23] M. Ghassemi et al., "A review of challenges and opportunities in machine learning for health," *AMIA Summits Transl. Sci. Proc.*, vol. 2020, pp. 191–200, 2020.

[24] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. 2008 IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, 2008, pp. 1322–1328.

[25] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Standardised metrics and methods for synthetic tabular data evaluation," Authorea Preprints, 2021. doi: 10.36227/techrxiv, 16610896.

[26] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.

[28] M. Jayabalan and M. E. Rana, "Anonymizing healthcare records: A study of privacy preserving data publishing techniques," *Adv. Sci. Lett.*, vol. 24, no. 3, pp. 1694–1697, 2018.

[29] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958.

[30] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: A review," *J. Imag.*, vol. 9, no. 4, 2023, Art. no. 81.

[31] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and Cox regression," *J. Mach. Learn. Res.*, vol. 20, pp. 1–30, 2019.

[32] D. H. Lee and S. N. Yoon, "Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges," *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, 2021, Art. no. 271.

[33] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[34] C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang, "VAEM: A deep generative model for heterogeneous mixed type data," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020, vol. 33, pp. 11237–11247.

[35] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinf.*, vol. 19, no. 6, pp. 1236–1246, 2018.

[36] R. Miotto et al., "Generative models for structured medical data: Challenges and opportunities," *JAMIA*, vol. 30, no. 6, pp. 1234–1245, 2023.

[37] R. M. Munshi, "Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction," *PLoS One*, vol. 19, no. 1, 2024, Art. no. e0296107.

[38] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano, "Synthetic data generation: State of the art in health care domain," *Comput. Sci. Rev.*, vol. 48, 2023, Art. no. 100546.

[39] OpenAI, "ChatGPT: Optimizing language models for dialogue," ChatGPT. OpenAI, 2022. Accessed: Sep. 5, 2025. [Online]. Available: chat.openai.com

[40] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[41] S. K. Pandey and R. R. Janghel, "Recent deep learning techniques, challenges and its applications for medical healthcare system: A review," *Neural Process. Lett.*, vol. 50, no. 2, pp. 1907–1935, 2019.

[42] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," in *Proc. VLDB Endow.*, Jun. 2018, vol. 11, no. 10, pp. 1071–1083. doi: 10.14778/3231751.3231757.

[43] B. Pereira et al., "The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes," *Nature Commun.*, vol. 7, May 2016, Art. no. 11479.

[44] V. Puri, S. Sachdeva, and P. Kaur, "Privacy preserving publication of relational and transaction data: Survey on the anonymization of patient data," *Comput. Sci. Rev.*, vol. 32, pp. 45–61, 2019.

[45] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[46] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[47] M. Schumacher et al., "Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group," *J. Clin. Oncol.*, vol. 12, no. 10, pp. 2086–2093, 1994.

[48] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to Learn*. Boston, MA, USA: Springer, 1998, pp. 3–17.

[49] J. W. Tukey, "The philosophy of multiple comparisons," *Stat. Sci.*, vol. 6, pp. 100–116, 1991.

[50] C. Umesh, M. Mahendra, S. Bej, O. Wolkenhauer, and M. Wolfien, "Challenges and applications in generative AI for clinical tabular data in physiology," *Pflügers Archiv- Eur. J. Physiol.*, vol. 477, pp. 531–542, 2024.

[51] M. J. V. der Laan, S. Dudoit, and K. S. Pollard, "Multiple testing. Part II. Step-down procedures for control of the family-wise error rate," *Statist. Appl. Genet. Mol. Biol.*, vol. 3, no. 1, Art. no. 14, 2004.

[52] W. G. Van Panhuis et al., "A systematic review of barriers to data sharing in public health," *BMC Public Health*, vol. 14 2014, Art. no. 1144.

[53] X. Wang et al., "MediTab: Scaling medical tabular data predictors via domain-informed pretraining," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 6062–6070.

[54] K. Woźnica, P. Wilczyński, and P. Biecek, "SeFNet: Bridging tabular datasets with semantic feature Nets," 2023, *arXiv:2306.11636*.

[55] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 7335–7345.

[56] S. Yang, F. Zhu, X. Ling, Q. Liu, and P. Zhao, "Intelligent health care: Applications of deep learning in computational medicine," *Front. Genet.*, vol. 12, 2021, Art. no. 607471.

[57] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," in *Proc. IEEE*, May 2021, vol. 109, no. 5, pp. 820–838. doi: 10.1109/JPROC.2021.3054390.