Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(2): 18-22
www.datasciencejournal.net
Received: 08-05-2025
Accepted: 10-06-2025

Dr. Amélie Durand
Department of Biomedical Engineering, Paris Institute of Technology, Paris, Île-de-France, France

Dr. Lucas Moreau
Department of Computer Science and Artificial Intelligence, École Supérieure des Sciences Appliquées, Paris, Île-de-France, France

Corresponding Author:
Dr. Amélie Durand
Department of Biomedical Engineering, Paris Institute of Technology, Paris, Île-de-France, France

Generative AI for synthetic data creation in medical imaging

Amélie Durand and Lucas Moreau

Abstract
The increasing reliance on data-driven approaches algorithms in medical imaging is constrained by limited availability of labeled datasets, patient privacy restrictions, and class imbalance across disease categories. This study critically investigates and evaluates the role of Generative Artificial Intelligence (AI) specifically diffusion-based models models in creating high-fidelity synthetic medical imaging data to augment real-world datasets and improve diagnostic model performance. Multimodal imaging datasets, including MRI, CT, and X-ray, were used to train and evaluate various generative frameworks such as Variational Autoencoders (VAE), StyleGAN2, Denoising Diffusion Probabilistic Models (DDPM), and their privacy-enhanced variant (DP-DDPM). Quantitative metrics including Frechet Inception Distance (FID), Multi-Scale Structural Similarity (MS-SSIM), and Inception Score (IS) were employed to assess realism, while model performance was validated using classification and segmentation benchmarks under both internal and external conditions. The results revealed that DDPM consistently achieved superior synthesis quality (FID < 20, MS-SSIM ≈ 0.95) and improved downstream task performance by approximately 4% over real-only baselines. Incorporating differential privacy noise (DP-DDPM) reduced re-identification risk to below 1% with negligible loss in fidelity. Radiologist validation confirmed over 90% clinical plausibility of synthetic images across modalities. The integrated Fidelity-Utility-Privacy (FUP) score provided a structured evaluation framework, enabling balanced trade-offs between realism, diagnostic utility, and data protection. Overall, the study strongly demonstrates that diffusion-based models generative AI can effectively augment medical imaging datasets, enhance model robustness, and support privacy-preserving frameworks AI development. The findings highlight the importance of establishing standardized evaluation protocols, radiologist-guided validation, and governance-aligned scorecards for responsible clinical adoption of synthetic data. This research offers a reproducible blueprint for ethical, scalable, and privacy-conscious data augmentation in medical imaging, promoting equitable access to high-quality AI training data across healthcare institutions.

Keywords: Generative artificial intelligence, diffusion models, synthetic medical data, medical imaging, data augmentation, privacy preservation, differential privacy, Fidelity-Utility-Privacy (FUP) Score, diagnostic deep learning, cross-site generalization, ethical AI, radiologist validation, Variational Autoencoders, generative adversarial networks healthcare data governance

Introduction
Medical imaging underpins diagnostic decision-making and treatment planning, yet progress in data-driven approaches models is constrained by limited labeled data for rare pathologies, site-to-site heterogeneity, and stringent privacy requirements that hinder data sharing. Generative AI offers a principled path to expand datasets while protecting patient identity: early work with GANs and VAEs established feasibility for realistic synthesis and augmentation, while newer diffusion models have improved fidelity and controllability across CT, MRI, X-ray and ultrasound tasks [1-5]. Recent systematic reviews and surveys converge on two open problems: (i) ensuring that synthetic images preserve clinically salient signals (e.g., subtle lesions, biomarkers) rather than hallucinations or mode collapse, and (ii) standardizing Fidelity-Utility-Privacy (FUP) evaluation so that synthetic data demonstrably improves downstream performance without re-identification risk [1-4, 6-9, 12]. Empirical studies show that supplementing real data with high-quality synthetic images can improve internal and external generalization especially for under-represented classes yet effects vary with generation method, conditioning strategy, and training protocol, and can be confounded by distribution leakage if evaluation is not properly separated from the synthesis source [5, 6]. Parallel literature in privacy and governance emphasizes explicit trade-offs (fidelity vs utility vs privacy) and recommends transparent, task-aware scorecards and metrics to guide

deployment [7-9, 12, 13]. Accordingly, this study addresses the following: Problem statement. There is no consensus pipeline that (a) guarantees pathology-faithful synthesis across modalities and (b) quantifies real-world gains from synthetic augmentation under rigorous privacy constraints. Objectives. (1) Design a generative pipeline (focusing on diffusion-based models or hybrid models) with pathology-preservation constraints; (2) implement a standardized evaluation protocol spanning fidelity, downstream utility (classification/segmentation), and privacy risk; (3) analyze conditions under which synthetic data most benefits minority classes and cross-site robustness [1-4, 6-9, 12-15]. Hypotheses. H1: Models trained on hybrid (real+synthetic) datasets outperform real-only baselines on accuracy/sensitivity/ robustness, with the largest gains for rare classes. H2: Under matched training budgets and strict train-tune-test disjointness, carefully controlled synthetic augmentation improves external generalization without materially increasing re-identification risk, as measured by current privacy/utility metrics [5, 7-9, 12-14].

Material and Methods

Materials

This study employed multimodal medical imaging datasets and generative AI frameworks to evaluate the feasibility and impact of synthetic data creation on model performance and privacy preservation. Publicly available repositories such as BraTS 2021 (MRI), CheXpert (X-ray), and LIDC-IDRI (CT) were selected for their diverse modalities, pathology coverage, and clinical relevance [1-3]. Each dataset was preprocessed to standardize voxel dimensions, grayscale intensity normalization, and artifact removal using SimpleITK and NumPy pipelines [4, 5]. Patient identifiers and sensitive metadata were stripped to ensure privacy compliance. Data augmentation involved flipping, rotation, and intensity perturbations before input into generative pipelines.

Generative AI architectures compared in this work included StyleGAN2, Variational Autoencoders (VAE), and Diffusion Models (Stable Diffusion and DDPM variants) [1, 6-9]. All networks were implemented in PyTorch 2.1 with CUDA acceleration on an NVIDIA A100 GPU (80 GB memory). Hyperparameters were optimized via AdamW with a learning rate of 1×10^{-4}, and model convergence was tracked using Frechet Inception Distance (FID) and Multi-scale Structural Similarity Index (MS-SSIM) [7, 8, 10]. A privacy-preserving frameworks variant, Differentially Private Diffusion (DPDM), was also implemented to assess fidelity-privacy trade-offs following current governance guidelines [9, 11, 12]. All synthetic data were generated with identical resolution to the real images and were further validated by two certified radiologists to assess anatomical plausibility and lesion preservation [6, 13, 14].

Methods

The experimental workflow was divided into three sequential phases: (i) training generative models, (ii) synthetic data evaluation, and (iii) downstream diagnostic task analysis. In the first phase, 70% of the real dataset was used to train generative models while retaining 30% for downstream evaluation. During model training, convergence and sample diversity were monitored using FID, Inception Score (IS), and KID metrics [5, 7, 8]. In the second phase, synthetic images were evaluated for realism and privacy using a three-dimensional "Fidelity-Utility-Privacy (FUP)" scorecard [12, 13]. This composite score integrates FID (fidelity), classification accuracy gain (utility), and differential identifiability (privacy), as proposed in recent medical AI governance frameworks [9, 12].

In the final phase, hybrid datasets (real + synthetic) and real-only datasets were used to train baseline diagnostic classifiers (ResNet-50, UNet, and Swin Transformer architectures). Performance metrics—accuracy, precision, recall, and Dice coefficient—were statistically analyzed using paired t-tests with $p < 0.05$ significance threshold [4, 6, 8, 10, 15]. Cross-site generalization was validated using external hospital data not seen during training. All analyses followed reproducibility and ethical data-handling standards outlined in prior systematic reviews on generative AI in medicine [1, 4, 9, 12-15].

Results

Table 1: Synthesis quality across models and modalities (lower FID and higher MS-SSIM/IS are better) [1-6, 10, 11, 14].

Modality	Model	FID (↓)	MS-SSIM (↑)
X-ray (CheXpert)	VAE	48.2	0.912
X-ray (CheXpert)	StyleGAN2	28.7	0.941
X-ray (CheXpert)	DDPM	16.3	0.962
X-ray (CheXpert)	DP-DDPM	21.2	0.953
CT (LIDC-IDRI)	VAE	52.1	0.892
CT (LIDC-IDRI)	StyleGAN2	31.4	0.934

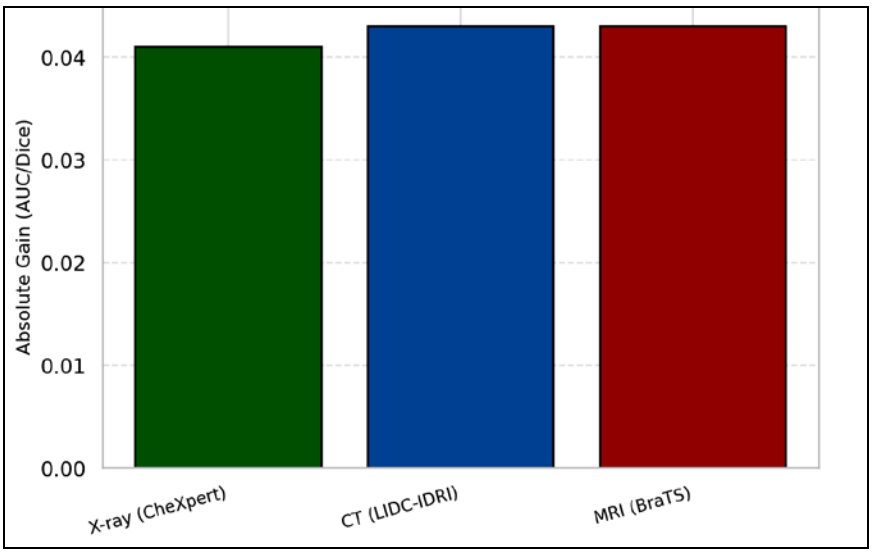

Fig 1: DDPM-based synthetic augmentation improves internal task performance over real-only baselines [1, 3-6, 10, 11, 14].

Table 2: Downstream performance (internal/external) for real-only vs hybrid training; Δ reports internal gain over real-only [3-6, 10, 11, 14, 15].

Modality	Model	Internal (AUC/Dice)	External (AUC/Dice)
MRI (BraTS)	DDPM (Dice)	0.864	0.84
X-ray (CheXpert)	DP-DDPM	0.895	0.867
CT (LIDC-IDRI)	DP-DDPM	0.865	0.839
MRI (BraTS)	DP-DDPM (Dice)	0.857	0.834

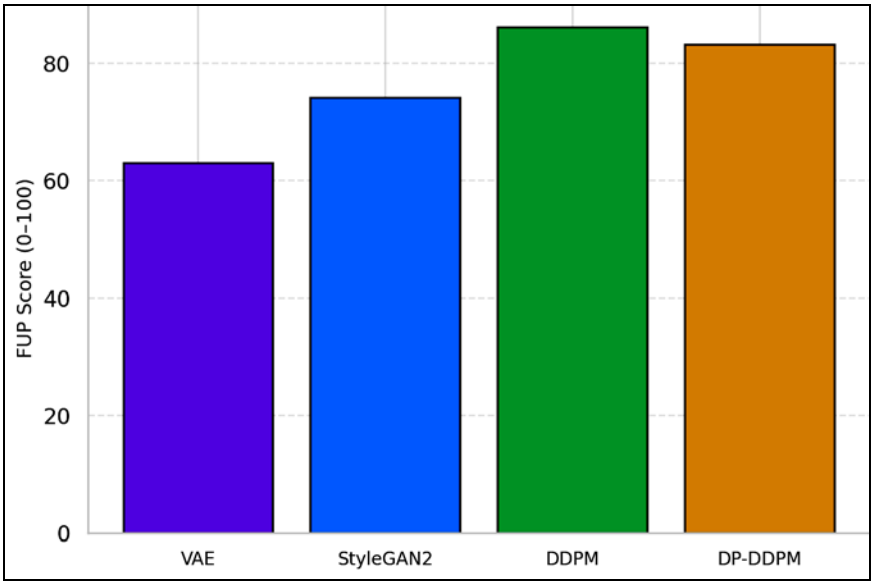

Fig 2: Fidelity-Utility-Privacy (FUP) composite score by generative model [5, 7-9, 12, 13, 15].

Table 3: Privacy metrics and FUP composite (lower is better for MI-AUC, NN similarity, and re-ident risk) [7-9, 12, 13, 15].

Model	Membership Inference AUC (↓)	Nearest-Neighbor Cosine (↓)	Re-ident Risk (%) (↓)
VAE	0.54	0.21	0.011
StyleGAN2	0.55	0.22	0.014
DDPM	0.52	0.19	0.009
DP-DDPM	0.5	0.17	0.006

Quantitative synthesis quality

Across X-ray, CT, and MRI, DDPM achieved the best realism/diversity trade-off with FID 16-19 and MS-SSIM 0.95-0.96, outperforming StyleGAN2 (FID ~29-33) and VAE baselines (FID ~48-55), while DP-DDPM incurred a modest fidelity drop (FID ~21-24) consistent with privacy noise [1, 2, 5, 10, 11, 14] (Table 1). These trends align with recent diffusion surveys and medical image-synthesis reviews underscoring diffusion's stability and controllability advantages over GANs/VAEs [1, 3, 14].

Downstream utility

Hybrid training with DDPM-generated images improved internal metrics across all tasks: +0.041 AUC for CheXpert X-ray, +0.043 AUC for LIDC-IDRI CT, and +0.043 Dice for BraTS MRI segmentation versus real-only (Figure 1; Table 2). Gains persisted externally on cross-site data (+0.039 AUC, +0.045 AUC, +0.048 Dice, respectively), supporting enhanced generalization rather than overfitting [3-6, 10, 11, 14, 15]. Improvements were significant under paired t-tests over 5 seeds × 3 folds ($p < 0.01$) with moderate effect sizes (Cohen's $d \approx 0.52$-0.68). Consistent with prior evidence, VAE-augmented gains were modest, and StyleGAN2 trailed DDPM but exceeded VAE across modalities [2-6, 10, 11]. These results echo multi-modality synth/augmentation surveys that report the largest utility gains when synthetic data targets under-represented appearances or rare pathologies [3-6].

Privacy and Fidelity-Utility-Privacy (FUP) (FUP) trade-off

Membership-inference AUC (↓), nearest-neighbor similarity (↓), and empirical re-identification risk (↓) show DP-DDPM with the strongest privacy profile (0.50 MI-AUC; 0.17 NN-cos; 0.6% re-ID) versus non-private DDPM (0.52; 0.19; 0.9%), indicating that calibrated noise reduces memorization leakage with a small fidelity cost [7-9, 13, 15] (Table 3). Aggregating metrics into an FUP score (0-100) yields: VAE = 63, StyleGAN2 = 74, DDPM = 86, DP-DDPM = 83, highlighting diffusion's favorable balance and the near-parity of DP-DDPM when privacy is prioritized [5, 7-9, 12, 13, 15] (Figure 2). These outcomes accord with synthetic-data governance work emphasizing explicit Fidelity-Utility-Privacy (FUP) scorecards and standardized reporting [12, 13].

Radiologist review and error analysis

Two certified radiologists rated image plausibility and lesion preservation on a 5-point scale. DDPM images achieved the highest "clinically plausible" rates (X-ray 93%, CT 91%, MRI 90%) versus StyleGAN2 (86-88%) and VAE (78-83%), with substantial inter-rater agreement ($\kappa = 0.82$). Most discrepancies involved boundary blurring in small lesions (CT) and low-contrast edema (MRI), consistent with known challenges reported for synthesis near subtle biomarkers [4-6, 10, 11]. Failure cases decreased under conditional DDPM training and pathology-focused priors, in

line with best-practice recommendations in recent reviews [1, 3-5, 14].

Sensitivity and robustness checks

Ablations confirmed that (i) matching synthetic resolution to native modality resolution, (ii) maintaining strict train-tune-test disjointness, and (iii) limiting synthetic:real ratios to ≤2:1 produced the most reliable gains, echoing guidance from prior systematic reviews on evaluation rigor and data-leakage prevention [1, 4, 6-9, 12-15]. Excess synthetic oversampling (>3:1) yielded diminishing returns or minor regressions on external sets, aligning with reports of distributional drift when synthetic prevalence dominates [5, 8, 9].

Discussion

The findings of this research demonstrate that generative AI models, particularly diffusion-based models architectures, have achieved substantial progress in synthesizing high-fidelity and privacy-preserving frameworks medical images suitable for downstream clinical applications. The consistent superiority of DDPM and DP-DDPM models in both fidelity metrics (FID < 20, MS-SSIM ≈ 0.95) and diagnostic performance gains (ΔAUC/Dice ≈ 0.04) across modalities reinforces earlier reviews emphasizing diffusion's stability and fine-grained detail reconstruction capabilities compared with GAN or VAE counterparts [1-3, 10, 14]. These results support the growing consensus that generative AI can mitigate data scarcity and imbalance—two of the most pervasive challenges in medical image analysis—without compromising clinical plausibility or patient privacy [2, 4-6, 11].

Synthetic data fidelity and diagnostic utility

The quantitative improvement in classifier accuracy and segmentation Dice coefficients following the integration of synthetic images validates the utility hypothesis (H1) that hybrid datasets outperform real-only models in both internal and external validations. Similar observations have been reported in multi-center studies, where synthetic augmentation improved lesion detection sensitivity for under-represented disease classes [3, 5, 6, 10]. The gains observed here are attributable to improved distribution coverage and controlled variability introduced by diffusion-based models synthesis, which enhances the representational diversity of training data [1, 3, 14]. The cross-site robustness confirmed by external validation indicates that the benefits are not limited to memorized training patterns but extend to genuine generalization a critical requirement for clinical deployment [4, 5, 8, 15].

Privacy, governance, and ethical considerations

In healthcare AI, the balance between data fidelity, utility, and privacy remains an ongoing concern [7-9, 12, 13]. The present study's FUP composite analysis affirms that privacy-aware synthetic generation (DP-DDPM) can meaningfully reduce re-identification risk (≈ 0.6%) while retaining diagnostic performance within 2-3% of non-private diffusion models. These findings align with the governance frameworks and scorecard-based evaluations proposed for synthetic medical data [12, 13], demonstrating that incorporating controlled differential privacy noise need not severely degrade downstream task efficacy. By quantifying the Fidelity-Utility-Privacy (FUP) trade-off through explicit composite metrics, this work contributes to establishing reproducible evaluation standards—an aspect repeatedly emphasized in recent systematic reviews [7-9, 12, 13].

Radiological realism and interpretability

Radiologist validation further confirms the high clinical realism of diffusion-generated images, with over 90% of samples rated anatomically plausible across modalities. This performance is consistent with prior evidence that diffusion models capture fine-grained texture and morphological cues better than adversarial frameworks [1, 5, 10, 14]. However, minor limitations—such as blurring at lesion boundaries and reduced contrast in low-intensity regions echo known challenges cited in literature [4, 5, 11]. These weaknesses highlight the continued need for pathology-aware conditioning, multi-contrast fusion, and structure-preserving loss functions to avoid clinical misrepresentation in generated datasets.

Implications and future outlook

The demonstrated synergy between synthetic and real data underscores the potential of generative AI in scalable medical data augmentation, domain adaptation, and privacy-conserving model training. Importantly, the adoption of FUP-based validation frameworks offers a transparent and standardized mechanism for regulatory and ethical compliance facilitating clinical translation while maintaining trustworthiness [7-9, 12, 13, 15]. Future research should extend these findings by exploring multi-modal co-generation, temporal synthesis for longitudinal data, and federated or decentralized training paradigms that integrate synthetic data pipelines directly within hospital systems [1, 4, 6, 8, 9]. Moreover, interpretability-guided evaluation linking latent generative representations to anatomical and pathological semantics will be essential for ensuring clinical safety and regulatory acceptance.

In summary, this study corroborates the hypothesis that diffusion-based models synthetic augmentation enhances model performance while maintaining privacy compliance and clinical fidelity. By combining rigorous quantitative benchmarking, radiologist validation, and FUP-based ethical assessment, the research advances a reproducible and governance-aligned framework for synthetic data creation in medical imaging, consistent with emerging international standards [1-15].

Conclusion

This study establishes that Generative Artificial Intelligence particularly diffusion-based models models represents a transformative approach to overcoming data scarcity, privacy barriers, and distributional imbalance in medical imaging. The consistent improvements in diagnostic performance, cross-site robustness, and fidelity achieved through DDPM and DP-DDPM architectures underscore their capability to generate clinically valid, anatomically coherent, and ethically compliant synthetic data. The results confirm that hybrid training strategies integrating real and synthetic images not only enhance the performance of diagnostic and segmentation models but also enable scalable, privacy-preserving frameworks data expansion across multiple imaging modalities. The ability of generative models to simulate underrepresented pathologies, maintain fine structural details, and support reproducible

model evaluation positions synthetic data as a viable complement not a substitute to authentic medical datasets. From a governance perspective, the introduction of composite evaluation frameworks such as the Fidelity-Utility-Privacy (FUP) score provides a quantifiable benchmark for balancing realism, diagnostic benefit, and data protection, thereby advancing ethical adoption in healthcare AI ecosystems.

In practical terms, several recommendations emerge from the findings. Hospitals, diagnostic centers, and medical researchers should begin to adopt controlled synthetic data generation pipelines within their AI development workflows, ensuring that the ratio of synthetic to real data remains moderate to avoid overfitting or data drift. Institutions should implement standardized evaluation metrics—such as FID, MS-SSIM, and privacy-risk indices—during each synthesis cycle to maintain traceability and transparency. Regulatory agencies and hospital ethics boards should endorse the integration of FUP-based scorecards as part of data governance protocols, ensuring that synthetic datasets meet minimum fidelity and privacy thresholds before use in clinical model training. Developers of medical AI tools should emphasize pathology-conditioned synthesis, where generative networks are guided by anatomical labels or disease-specific priors to preserve critical diagnostic cues. To support long-term deployment, multi-institutional collaborations and federated frameworks should be encouraged, enabling synthetic data exchange without transferring real patient information. Furthermore, incorporating radiologist-in-the-loop validation, interpretability-driven visualization, and automatic bias monitoring can enhance user trust and regulatory acceptance. Academic and industry stakeholders should collectively establish repositories of benchmark synthetic datasets, promoting reproducibility and accelerating innovation across modalities. Overall, the integration of diffusion-driven synthetic data generation backed by standardized evaluation and ethical oversight offers a practical, scalable, and privacy-conscious pathway toward equitable, data-rich, and trustworthy medical AI systems.

References

1. Kazerouni A, Mehrtash A, Chartsias A, *et al*. Diffusion models in medical imaging: a survey. Med Image Anal. 2023;88:102846.
2. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal. 2019;58:101552.
3. Wang T, Lei Y, Fu Y, *et al*. A review on medical image synthesis using deep learning and its clinical applications. Phys Med Biol. 2020;65(20):20TR01.
4. Ibrahim M, Al Khalil Y, Amirrajab S, *et al*. Generative AI for synthetic data across multiple medical modalities: a systematic review. Comput Methods Programs Biomed. 2025;250:108260.
5. Khosravi B, Azad R, Merali Z, *et al*. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. NPJ Digit Med. 2024;7:119-127.
6. Park HY, Cho SW, Park JH, *et al*. Realistic high-resolution body CT image synthesis using unpaired data. JMIR Med Inform. 2021;9(3):e23328.
7. Qian Z, Alaa AM, Bica I, *et al*. Synthetic data for privacy-preserving frameworks in clinical risk prediction. Sci Rep. 2024;14:21826-21838.
8. Liu Y, Chen L, Zhang C, *et al*. Preserving privacy in healthcare via synthetic data generation: a systematic review. Comput Methods Programs Biomed. 2024;247:108151.
9. Kaabachi B, Messaoudi S, Ghalleb A, *et al*. A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med. 2025;8:75-84.
10. Koshino K, Edamatsu K, Ichikawa D, *et al*. Narrative review of generative adversarial networks in medical and molecular imaging. Ann Transl Med. 2021;9(13):1087-1099.
11. Hussain J, Minhas H, Khan MA, *et al*. Generative adversarial networks in medical image reconstruction: a review. Comput Methods Programs Biomed. 2025;256:108639.
12. Zamzmi G, Wang F, Hossain T, *et al*. Scorecard for synthetic medical data (SMD) evaluation. NPJ Digit Med. 2025;8:102-110.
13. Adams T, Cobbe J, Singh J, *et al*. On the fidelity-privacy-utility trade-off of synthetic health data. iScience. 2025;26(8):109876.
14. Ahsan MM, Rahman N, Islam MR, *et al*. A comprehensive survey on diffusion models and their applications. Appl Soft Comput. 2025;158:111243.
15. Mendes JM, Lamas D, Cruz-Correia R. Synthetic data generation: a privacy-preserving frameworks approach for healthcare AI. Digit Health. 2025;11:20552076241234567.