International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 4, PP 225-231, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I4P126

Original Article

Synthetic Data Generation for Validation of Clinical Research Software

Rohit Singh Raja
Principal Software Quality Engineer.

Received On: 19/10/2025 ***Revised On: 20/11/2025*** ***Accepted On: 03/12/2025*** ***Published On: 13/12/2025***

Abstract: *Validation of clinical research software is limited by the lack of realistic, privacy-safe datasets capable of exercising complex protocol and workflow logic. This study proposes a hybrid synthetic-data generation framework that combines deterministic clinical simulation, deep generative modeling, and differential privacy to create statistically faithful and audit-traceable datasets tailored for DCT and eCOA validation. In a Phase-II–scale evaluation, the approach achieved 38% higher defect detection, JS divergence = 0.054, and membership-inference AUC ≈ 0.52, demonstrating that synthetic data can support scalable, privacy-preserving, and empirically rigorous validation of regulated clinical-research platforms.*

Keywords: *Synthetic Data, Data Generation, Clinical Research Software, Software Validation, Data Simulation, Clinical Data Modeling, Electronic Health Records (EHR) Simulation, Patient Data Privacy, Regulatory Compliance.*

1. Introduction

Clinical research software development platforms utilizing EDC, eCOA, RTSM and DCTs, are subject to the same requirements for robust testing as all clinical research software; because errors may potentially affect the participants' safety, study integrity and regulatory compliance. However, despite the potential impact of errors in clinical research software, the validation process continues to utilize small, manual, or partially de-identified datasets which lack the ability to accurately depict the complex aspects of clinical trials such as varying visit window times, missing values, dropout of device data, and violations of the protocol. Therefore, critical paths in the validation process will continue to go untested, reducing the detection of defects in the software.

Regulatory agencies such as the FDA and EMA advocate for lifecycle verification, traceability and Risk-Based Quality Management (RBQM); however, they provide no practical assistance to achieve realistic, patient-like datasets that have been de-identified in a way that preserves confidentiality and is also permissible for distribution across engineering and QA teams. Consequently, this creates a basic operational problem: current clinical research software must be tested with patterns of data that are not allowed to be retained by the software for use in testing. Synthetic patient data represents one possible solution. Recent advances in deterministic clinical simulation, generative modeling (GANs, VAEs, diffusion models), and differential privacy allow for the generation of longitudinal patient-like datasets that maintain the statistical properties of real patient data and eliminate the need to expose real patient data. Synthetic patient data can represent adverse event cascades, diary rhythms, protocol violations and many other edge cases in clinical trials that are typically not found in typical validation datasets and expand the scope of validation testing.

This paper proposes a hybrid approach for generating synthetic data using a combination of deterministic clinical simulation, deep generative models and differential privacy to create statistically valid, confidential and auditable patient-like datasets for validating DCT and eCOA systems. Additionally, this paper presents a QA integration layer to convert synthetic patient records to executable test artifacts, and an evaluation methodology that assesses the validity of the generated synthetic patient data, the potential for leaking confidential information from the synthetic patient data, and the utility of the synthetic patient data for validation testing. Together, these three components provide a scalable, regulated and self-sustaining basis for automated validation in clinical research software.

2. Literature Review

The use of artificial data is being studied in the context of three areas of research: synthetic electronic health record (EHR) creation, privacy-protective transformation of data, and validation of generated data within regulatory clinical-research systems. The first simulation platforms (SyntheaTM) showed that deterministic, rule-based systems could create clinical realistic longitudinal profiles; however, their limited variability in timing, interaction between comorbidities and missing data patterns limit their ability to validate software. The development of deterministic rule-based systems was replaced with neural generative models. Neural generative models have been able to simulate high dimensional patient records using adversarial training (medGAN and its successors) and have extended the prior work by including loss functions based on the Wasserstein distance, generative models that incorporate temporal relationships, and graph-based representations of comorbidities to better model the distributional characteristics and longitudinal nature of patient data. Diffusion-based models have recently improved the realism of irregular time intervals, and therefore, they have the potential to accurately reconstruct diary behaviors and protocol dependent timing of visits.

At the same time, there has been an accumulation of literature showing that synthetic data sets are not inherently anonymous; generative models can reveal identifiable information in the training data. Differential Privacy (DP) has emerged as a way to mitigate this risk by bounding the memorization of sensitive information through noisy optimization. Although DP-GAN and related frameworks provide strong theoretical guarantees, they rarely contain audit ready documentation, provenance and traceability required for HIPAA Expert Determination or for compliance with HHS guidelines for de-identifying data. This limitation is especially detrimental for validating DCTs and eCOAs, because synthetic data must support 21 CFR Part 11 compliant audit trails, reproducibility and dissemination of data across engineering and quality assurance teams, while maintaining confidentiality.

Clinical-software validation literature provides additional requirements that are not addressed by most synthetic data research. In regulated environments, traceability, reproducibility and risk-based validation consistent with ICH E6(R3) and RBQM principles are emphasized. Most commonly, defects occur at the edges of clinical software validation due to edge case scenarios, such as adverse events occurring outside of expected windows, visits occurring outside of expected windows, gaps in telemetry from devices and back filling diaries, all of which are infrequently captured in traditional or machine-learning oriented synthetic data. Synthetic data is currently evaluated almost exclusively using ML utility metrics (i.e. predictive accuracy), which do not evaluate software validation utility (i.e. coverage of logical paths and stress testing of protocol rules). This mismatch creates a significant gap in methodology between academia and the requirements of regulated clinical-research platforms.

Finally, ethical and governance considerations create additional limitations. Generative models may exacerbate biases contained within the underlying statistical priors, particularly if demographic subgroups are underrepresented. Recent regulatory commentary has highlighted the need for fairness assessments, representativeness evaluations, and lineage documentation to ensure that generative models are used responsibly. Furthermore, these requirements go beyond data privacy to include transparency, explainability, and governance, which are appropriate for GxP regulated environments.

Table 2.1 illustrates the divergence between current synthetic-data capabilities and clinical-software validation requirements. Current research is focused on replicating distributions, adding noise for privacy, and evaluating utility metrics based on ML performance; regulated validation requires order-fidelity of events, tracing anomalies, coverage of protocol-logical pathways, justification of privacy consistent with CFR standards and incorporation into automated Continuous Integration / Continuous Deployment (CI/CD) workflows. This study will address this gap by developing a hybrid generative/validation framework that is designed to meet the clinical-quality and regulatory-compliance requirements for clinical QA.

Table 1: Comparison of Existing Synthetic-Data Techniques vs. Requirements for Clinical Software Validation

Dimension	**Existing Synthetic Data Research**	**Clinical Research Software Validation Needs (DCT/eCOA)**	**Gap**
Statistical Fidelity	GANs, VAEs, graph models replicate distributions [3], [4], [6], [12].	Distributional accuracy + event-order fidelity + missingness patterns.	Temporal and structural fidelity underexplored.
Privacy & HIPAA Compliance	DP-GANs; risk reviews [5], [8], [12].	Formal de-identification, CFR compliance, audit justifications.	Need for audit-ready privacy justification.
Validation Utility	Utility measured via ML performance.	Needs edge-case generation, stress-testing, logic-path coverage.	ML utility ≠ QA software utility.
Regulatory Alignment	Mostly academic; limited to privacy.	Must satisfy 21 CFR Part 11, ICH E6(R3), RBQM, QbD.	No regulatory-aligned pipelines exist.
Automation	Few works on integration into testing workflows.	Required for automated regression, CI/CD validation.	Lack of integration architecture.

3. Methodology

The methodology is divided into four parts: (1) generating synthetic data, (2) maintaining privacy while providing auditability, (3) integrating quality assurance (QA) into workflow, and (4) developing an evaluation framework. In addition, the methodology includes design elements (traceability, reproducibility, and privacy) that meet DCT/eCOA validation requirements as well as those of regulatory compliance (21 CFR Part 11, HIPAA, and ICH E6(R3)).

3.1. Synthetic Data Generation Workflow

The hybrid generation workflow utilizes a combination of deterministic protocol simulation and deep generative modeling to generate clinically plausible, variable, and audit-traceable synthetic patient trajectory data.

3.1.1. Deterministic Clinical Simulator

This simulator provides a modular rule-based method of encoding the protocol's components (eligibility criteria, randomization rules, visit windows, dosing schedules, and AE/SAE probability curves) using clinical state machines. The use of clinical state machines allows for the generation of trajectories that are consistent with medically plausible event ordering and medically relevant protocol dependency logic required for the validation of DCT/eCOA systems.

3.1.2. Tabular GAN (medBGAN variant)

Tabular GANs (Generative Adversarial Networks) model the joint distribution of multiple types of demographic data (demographics), laboratory data (labs), symptom data (symptoms), patient reported outcome (PRO) data (entries), and device signal data. Tabular GANs also employ conditional sampling and anti-mode collapse regularization to maintain rare but clinically significant patterns (e.g., severe AEs, atypical lab values) that are important for the QA-stress testing and logic path coverage.

3.1.3. Continuous Time Diffusion Model

Diffusion models generate the temporal structure of the synthetic data (i.e., inter-visit interval, diary rhythm, telemetry burst/gap pattern, and missingness behavior). Temporal structure is a key element for validating eCOA logic, diary window enforcement, and time dependent protocol checks.

3.1.4. Anomaly Injection Layer

In order to validate regulatory compliant systems beyond "happy-path" scenarios, the anomaly injection layer injects controlled deviations into the synthetic data, including timestamp drift, diary backfill attempts, misaligned eSign events, missing entries, and device dropout cascade, each carrying metadata lineage for auditability. This enables comprehensive testing of exception handling logic in regulated systems.

3.2. Privacy Preservation and Auditability Differential Privacy (DP)

GAN training uses DP-SGD with gradient clipping and calibrated noise. Privacy accounting yields **$\varepsilon \approx 2.4$, $\delta = 1e\text{-}5$**, balancing anonymity with statistical fidelity.

3.2.1. Post-Generation Privacy Audits

To satisfy HIPAA Expert Determination and support internal privacy governance, the following audits are conducted:

- Membership-Inference Attack (MIA) AUC Evaluates memorization; near-0.5 AUC indicates low leakage.
- Distance to Closest Record (DCR) Ensures synthetic samples are sufficiently dissimilar from nearest real profiles.
- Attribute-Inference Tests Detect potential leakage of sensitive demographics or clinical attributes.
- Rare-Case Leakage Analysis Ensures low-prevalence clinical combinations are not reproduced verbatim.

3.2.2. Traceability and Regulatory Alignment

Each synthetic record includes:

- cryptographic hashes
- generator versioning and DP parameters
- scenario and anomaly lineage
- time stamped audit-trail entries (21 CFR Part 11–aligned)

This metadata enables deterministic replay, supports GxP audits, and ensures full traceability across validation pipelines.

3.3. QA Simulation and Integration Layer

This layer transforms synthetic patients into executable validation artifacts that integrate directly into DCT/eCOA QA workflows and CI/CD pipelines.

3.3.1. Artifact Packaging

Synthetic trajectories are converted into:

- JSON/XML visit logs
- PRO/diary time-series
- device telemetry sequences
- audit-trail events (user actions, timestamps, eSignatures)

Test runners ingest these artifacts to evaluate system behavior under controlled and reproducible conditions.

3.3.2. Scenario Bundles and Automation

Patients are grouped into thematic bundles to exercise specific validation domains:

- protocol-deviation scenarios
- high-missingness diaries
- AE-escalation cascades
- mixed telemetry modalities
- timestamp inconsistency bundles

These bundles support automated regression testing and RBQM-aligned risk-focused validation.

3.3.3. Expected-Outcome Rules

Each scenario includes a rule file defining expected system behavior (e.g., "visit must be marked out-of-window," "AE severity must escalate workflow," "eSign validation must fail"). This enables automated pass/fail scoring without manual review, accelerating test cycles and improving repeatability.

3.4. Evaluation Metrics

The evaluation framework measures fidelity, privacy, and validation utility, mapped directly to regulatory and QA needs.

3.4.1. Fidelity Metrics

- Kolmogorov–Smirnov (KS) Distance Captures tail differences essential for detecting unrealistic extreme labs/vitals.

- Jensen–Shannon (JS) DivergenceAssesses distributional similarity for categorical variables such as AE types.
- Correlation-Matrix Similarity Evaluates multivariate dependencies important for AE–lab–vital interactions.
- Temporal Alignment Score Measures adherence to protocol windows critical for validating visit logic and diary enforcement.

3.4.2. Privacy Metrics

- MIA AUC (memorization resistance)
- DCR thresholds (replicate detection)
- DP parameter verification (ε, δ compliance)
- rare-case leakage checks

3.4.3. QA Utility Metrics

- Defect-Detection Rate – primary validation effectiveness measure
- Scenario Coverage Index (SCI) – breadth of logic-path and protocol coverage
- Rule-Violation Recall – system's ability to catch protocol deviations
- End-to-End Workflow Completion – robustness under stress scenarios
- Automated Test-Script Generation Success – CI/CD readiness

3.4.4. Regulatory Mapping

Table 2: Compliance Requirements and Metric Mapping

Requirement	Metric Mapping
RBQM risk detection	Fidelity metrics (KS, JS, correlation, temporal)
HIPAA/HHS de-identification	DP parameters, MIA AUC, DCR
21 CFR Part 11 / GxP	Traceability metadata, audit trails
ICH E6(R3)	Scenario coverage, edge-case validation

Algorithm 1. Synthetic-Data QA Generation Workflow
Input: protocol *P*, statistical profiles *D*, DP budget
Output: synthetic dataset *S*, audit artifacts *A*

```
1. Build protocol-driven state machines from $P$.
2. For each simulated patient:
   a. Generate baseline trajectory using a deterministic simulator.
   b. Sample tabular features via DP-GAN conditioned on trajectory.
   c. Generate temporal sequences via diffusion model.
   d. Inject anomalies based on scenario design.
   e. Package QA artifacts (JSON/XML, telemetry, audit logs).
3. Run privacy audits (MIA, DCR, attribute-inference, rare-case checks).
4. Compute evaluation metrics (KS, JS, correlation similarity, temporal alignment).
5. Export $S$ and $A$ for automated QA and CI/CD pipelines.
```

3.4.5. Computational Setup

Models were trained on an NVIDIA A100 GPU, 32-core CPU, and 128 GB RAM. DP-GAN training increased compute cost by ~1.5× relative to non-DP GANs. Diffusion models converged in 12–18 hours. Generating 12,000 synthetic patients (~2.8M timestamped records) required ~22 minutes. Additional reproducibility details may be included in an appendix per journal requirements.

4. Results & Analysis

This study tested the quality, privacy, and usability of a new method called "hybrid" of generating synthetic medical records through both completely synthetic and simulated methods.

4.1. Statistically Valid

The hybrid model combining deterministic simulation, GAN-based feature modeling and diffusion driven temporal synthesis produced very good statistically valid results on all analyzed parameters. For continuous variables (lab tests, vital signs, Patient Reported Outcome (PRO) scores), the average Kolmogorov-Smirnov distance of KS_mean = 0.073 was obtained; this means that the synthetic distributions were nearly identical to the reference distributions. The categorical variables had a JS-divergence of 0.054, which indicates that they had realistic frequencies. The multivariable structure was also well maintained as evidenced by a correlation matrix similarity of 0.91. This is particularly important for retaining clinical interdependencies between variables. The temporal realism of the models was also very good with a temporal alignment score of 0.88 and 93.4% accurate visit window. These two metrics indicate that the time frames of the synthetic visit trajectories were medically reasonable, which is a critical aspect of eCOA and DCT workflow validation.

4.2. Privacy Protection

Minimal risk of re-identification existed based on privacy evaluations. The membership inference attack on the shadow model resulted in an area under the receiver operator characteristic curve (AUC) of 0.52, which represents random guessing. The minimum distance to closest record exceeded 0.41, representing the threshold acceptable for synthetic datasets. The differential privacy accounting provided an epsilon value of 2.4 (delta = 10^-5), providing an upper bound on memorization risk while preserving statistical utility. Collectively, these values represent compliance with HIPAA expert determination standards, and further indicate that the fidelity gains in this work did not occur at the cost of privacy.

4.3. Validation Utility in DCT/ECOA QA Workflow

Inclusion of the synthetic datasets in automated validation workflows significantly improved QA efficiency. Over 500 validation executions, the inclusion of the synthetic datasets improved defective-detection of the artificially created faulty patient-traces by 38% over use of the hand-authored datasets. The greatest improvements occurred with respect to defects involving complex logic inconsistencies, such as visit-window misaligned, timestamps irregularly occurring, or conflicts regarding the severity of AE/PROs. The scenario coverage, defined as the proportion of possible scenarios tested, improved from 0.52 to 0.87, indicating that the synthetic datasets permitted testing of a greater variety of clinically unusual edge-cases than would be typical of author-created test data.

4.4. Benchmarking Against a GAN-Only Generator

Comparison with a baseline medBGAN-only generator demonstrated measurable gains across fidelity, privacy, and QA utility. The hybrid model achieved:

- KS distance: **0.112 → 0.073**
- JS divergence: **0.088 → 0.054**
- Correlation similarity: **0.82 → 0.91**
- Temporal alignment: **0.71 → 0.88**
- Privacy AUC: **0.61 → 0.52**

Correspondingly, QA metrics improved:

- Scenario Coverage Index: **0.71 → 0.87**
- Defect-detection rate: **0.68 → 0.82**

These results indicate that combining diffusion modeling and DP mechanisms with a deterministic simulator produces more realistic, temporally coherent, and privacy-robust datasets than GAN-only approaches.

4.5. Operational Feasibility Testing

Testing of operational feasibility was conducted using a focused audit trial. This consisted of generating 500 synthetic patient audit logs and injecting 12 different anomalies to evaluate if the system was ready for use with real world validation workflow processes. All twelve were identified by the system except one that had a very low frequency permutation of AE-codes; this resulted in 11/12 or 91.7% identification of all anomalous activity which included clinical and operational relevant discrepancies between diary date/time entries and actual visit date/time entries.

The hybrid synthetic data methodology has shown:

- Statistical fidelity for both distributions, correlation and time-based structures at a high level
- Low levels of privacy leakage due to differential privacy accounting and adversarial testing
- Substantial increases in quality assurance utility as evidenced by 38% more defects found than baseline methods
- Better performance compared to GAN only baselines
- Operational feasibility and potential for integration into automated DCT/eCOA validation systems.

These results support the framework's suitability as a scalable, privacy-preserving, and regulation-aligned tool for validating clinical research software.

Table 3: Comparison of Synthetic-Data Fidelity and QA Utility

Metric	GAN Only	Hybrid Model (GAN + Diffusion + DP)
KS Distance (↓)	0.112	0.073
JS Divergence (↓)	0.088	0.054
Correlation Similarity (↑)	0.82	0.91
Temporal Alignment (↑)	0.71	0.88
Privacy Attack AUC (≈0.5 ideal)	0.61	0.52
Scenario Coverage Index (↑)	0.71	0.87
Defect Detection Rate (↑)	0.68	0.82

The hybrid approach significantly outperforms single-model generation especially in temporal fidelity and QA utility metrics.

5. Discussion and Future Scope

These results show that hybrid architectures of synthetic clinical data with a combination of deterministic clinical simulation, deep generative models and differential privacy provide a scalable and regulatory compliant base for validating clinical research software. The improved temporal fidelity and increased breadth of coverage of scenarios and defects shown here are also consistent with many of the newer regulatory expectations under ICH E6(R3) for Clinical Trials Data Integrity, Quality by Design (QbD) and Risk-Based Quality Management (RBQM). In addition to these benefits, because synthetic datasets can be engineered to include specific CtQs such as incorrect visit windows, adverse events that trigger additional events, or incorrectly reported diary entries, etc., they allow for proactively identifying risks to product quality in a way that would otherwise require extensive testing of production systems to ensure defects do not leak into those systems.

In addition to providing the ability to perform proactive, risk-based testing on clinical trial software, the framework provides the traceability and auditability features necessary to meet the requirements of 21 CFR Part 11 and more broadly, the GxP regulations. A key aspect of the framework is the inclusion of provenance information in each generated synthetic patient record which includes information about how it was generated, what anomalies were included in it and how it was transformed to create the final version. This allows for deterministic replay and reproducibility of the testing process that cannot be done with manually constructed test datasets.

While differential privacy has provided an effective method to protect the privacy of individuals in the synthetic data, it is still important to have governance practices in place to ensure that there are no unintended biases introduced into the synthetic data from the source statistical distributions, particularly if the source data does not adequately capture all demographics. Therefore, routine fairness audits and demographic coverage reviews should occur prior to deploying synthetic data sets. Additionally, while generative models used to generate synthetic data are difficult to interpret, documentation and explanation of the models will be needed including model cards, training data limitations, and human review processes will need to be established for high-risk clinical use cases such as SAE workflow logic, adjudication patterns, and protocol-driven event sequence.

Looking ahead, several extensions could amplify the framework's clinical and regulatory utility.

- CI/CD-Integrated Continuous Validation: The inclusion of synthetic-data engines in continuous-integration/continuous-deployment (CI/CD) automation pipelines enables to test regressions on each software release; providing the ability to monitor compliance with protocol-adherence logic and validate changes in validation drift continuously.
- New QA Metrics: Metrics such as Rule-Coverage Density, Protocol Deviation Simulation Rates and Validation Path Completeness (VPC), can provide more specific quantification of how synthetic-datasets test clinical-decision paths and are superior to the standard metrics currently used to measure distributional-fidelity.
- End-to-End Workflow Testing Across Multiple Systems: Through synchronized synthetic-patient timelines that span all systems within the EDC, RTSM, eCOA, eConsent and Device Telemetry system, the framework supports comprehensive end-to-end workflow validation, which is one of the areas where current QA practices are weak.
- Explainability Enabled Generative Models: With increasing requirements from regulators for transparency, the addition of explanation tools (i.e., SHAP-based feature influence vector, generation-trace visualization, etc.) may assist in documenting predictable behavior and operational limits of synthetic-data engines in use within regulated environments.

The hybrid synthetic-data framework provides a practical, privacy-respecting, and auditable method for enhancing the validation of clinical research software using real patients. Synthetic patient generation represents a scalable validation asset that provides a base for more sophisticated, automated, and explainable QA architectures by improving defect detection and expanding coverage of scenarios.

6. Conclusion

This study has developed an integrated, regulation compliant approach to generate synthetic patient data to be used for validating clinical research software. This is achieved through a combination of deterministic clinical simulation, generation of features using generative adversarial networks (GANs) and generation of temporal models using diffusion processes. Furthermore, the framework integrates differential privacy into these elements and therefore generates realistic synthetic data sets with full provenance which are safe under HIPAA and usable in GxP environments. The empirical evaluation of this framework has demonstrated its high degree of statistical fidelity to the original data set, maintained the temporal relationships between variables and the correlations among them, while also maintaining low privacy risks. In addition, membership inference attacks were performed at random levels. The use of this type of synthetic data set in the context of automated DCT/ eCOA QA pipelines resulted in a greater number of scenarios being tested and a 38% increase in defects detected and therefore demonstrated a significant amount of practical application.

The auditability and traceability capabilities provided in the framework align with regulatory requirements including ICH E6(R3), QbD, RBQM, and 21 CFR Part 11 to provide a reproducible and privacy compliant basis for the verification of clinical research software. Synthetic data sets serve as both a safe substitute for actual data and as engineered validation assets that will allow for continued testing, improved risk detection and expanded coverage of protocol sensitive workflows. Overall, the study demonstrates that hybrid synthetic data can be a central component in modernizing validation practices throughout the clinical research technology community.

References

1. Gonçalves, C. Ray, and C. Rusu, “Generation and evaluation of synthetic patient data,” *BMC Medical Research Methodology*, vol. 20, no. 108, 2020, doi: 10.1186/s12874-020-00977-1.
2. J. Walonoski *et al.*, “Synthea™: A synthetic patient generator for benchmarking health IT tools,” *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 230–238, 2018, doi: 10.1093/jamia/ocx079.
3. M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, “Synthesizing electronic health records using improved generative adversarial networks,” *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019, doi: 10.1093/jamia/ocy142.
4. E. Choi *et al.*, “Generating multi-label discrete patient records using generative adversarial networks,” in
5. *Proceedings of the Machine Learning for Healthcare Conference*, vol. 68, PMLR, 2017, pp. 286–305.
6. Torfi, E. A. Fox, and C. K. Reddy, “Differentially private synthetic medical data generation using convolutional GANs,” *Information Sciences*, vol. 586, pp. 485–500, 2022.
7. G. Nikolentzos *et al.*, “Synthetic electronic health records generated with graph-based deep generative models,”
8. *NPJ Digital Medicine*, 2023.
9. K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. Sebastopol, CA, USA: O’Reilly Media, 2013.
10. K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, “A systematic review of re-identification attacks on health

data," *PLoS ONE*, vol. 6, no. 12, e28071, 2011, doi: 10.1371/journal.pone.0028071.

11. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLoS Digital Health*, vol. 2, no. 1, e0000082, 2023, doi: 10.1371/journal.pdig.0000082.
12. M. Beigi *et al.*, "Simulants: Synthetic clinical trial data via subject-level simulation," *Contemporary Clinical Trials Communications*, 2023, doi: 10.1016/j.conctc.2023.101182.
13. K. El Emam, "Utility metrics for evaluating synthetic health data generation methods," *JMIR Medical Informatics*, 2022, doi: 10.2196/38143.
14. A. Naseer *et al.*, "ScoEHR: Synthetic electronic health records generation with continuous-time diffusion models," in *Proceedings of Machine Learning and Systems*, 2023.
15. U.S. Food and Drug Administration, *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*, FDA, 2021.
16. European Medicines Agency, *Reflection Paper on the Use of Artificial Intelligence in the Medicinal Product Lifecycle*, EMA/36932/2023, 2023.