Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 6, 2026

FIDELITY AND UTILITY OF SYNTHETIC TABULAR HEALTH DATA: A CROSS-PARADIGM BENCHMARKING OF A CORRELATION-PRESERVING STATISTICAL PIPELINE AND A CONDITIONAL GENERATIVE ADVERSARIAL NETWORK

[1]Ishrat Fatima, [2]Najwa Liaqat, [3]Mahnoor Saeed
[1]*Visiting Lecturer, Department of Statistics, University of Sargodha*
[2]*Department of Statistics, COMSATS University Islamabad (CUI) Lahore Campus*
[3]*Department of Statistics, Comsats University Islamabad, Lahore Campus*
[1]ishratf422@gmail.com, [2]najwa.liaqat.nl@gmail.com, [3]mahnoorsaeed925@gmail.com

DOI: https://doi.org/10.5281/zenodo.20727134

Keywords
Synthetic data generation; generative adversarial networks; CTGAN; statistical simulation; tabular health data; Modgo; correlation preservation; downstream utility; preprocessing sensitivity

Article History

Received on 10 May, 2026
Accepted on 14 June, 2026
Published on 17 June, 2026

Corresponding Author: *

Abstract

Background: The generation of synthetic tabular health data has emerged as a pivotal methodology for statistical simulation, privacy preservation, and methodological validation. While classical statistical approaches offer interpretable, correlation-preserving synthesis through explicit parametric modeling, generative adversarial networks (GANs) provide a flexible, non-parametric alternative that learns the full joint distribution directly from empirical data. Despite the rapidly expanding application of both paradigms, a rigorous, cross-disciplinary empirical comparison tailored to the specific challenges of mixed-type health datasets with small-to-moderate sample sizes remains absent. Aims & Objective: This study evaluates whether a GAN-based approach (CTGAN) offers demonstrable advantages over a modern classical statistical pipeline for synthetic health data generation, and under which practical conditions each method excels. Methodology: Six publicly available tabular health datasets were selected, encompassing continuous, binary, and time-to-event outcomes with mixed continuous and categorical predictors. The statistical pipeline comprised the Modgo method for covariate matrix simulation, followed by parametric model-based outcome generation (linear, logistic, or Cox models). The CTGAN architecture was trained for 1000 epochs. Synthetic data quality was assessed through quantitative distributional similarity metrics (moments, Kullback–Leibler divergence, Kolmogorov–Smirnov statistic, Jensen–Shannon distance, Wasserstein distance), categorical concordance, pairwise correlation recovery, and downstream predictive utility measured via mean squared error, balanced accuracy, and concordance index. Sensitivity to variable scaling and training epoch selection was systematically examined. Results & Findings: Both methods successfully reproduced univariate distributions and joint dependence structures across all datasets. The statistical approach demonstrated superior fidelity in recovering pairwise linear correlations, attributable to its explicit correlation matrix estimation. Consequently, it yielded marginally higher prediction performance for binary classification (balanced accuracy 0.85 vs. 0.78) and survival analysis (C-index 0.71 vs. 0.65). The CTGAN approach achieved lower mean squared error in linear regression (1.02 vs. 2.82) and exhibited pronounced robustness to feature scaling, in contrast to the statistical method which failed catastrophically without prior standardization. CTGAN-

generated densities occasionally displayed discretized artifacts, particularly in small samples. Both methods performed comparably in categorical distribution recovery and overall downstream utility. Conclusions: Classical statistical and GAN-based synthetic data generation methods are both viable tools for health simulation studies, each with distinct operational strengths. The statistical pipeline is recommended when the precise reproduction of known correlation structures or parametric outcome relationships is critical, provided that rigorous data preprocessing is undertaken. The CTGAN approach is preferable for datasets with complex multimodality, heavy tails, or when minimal preprocessing and automated implementation are desired. These findings sanction the principled integration of AI-based generators into the statistical simulation repertoire and underscore the importance of context-dependent method selection. Future research should prioritize the incorporation of privacy preservation mechanisms and the development of hybrid architectures that couple the interpretability of statistical models with the distributional flexibility of deep generative networks.

1. Introduction

Simulation constitutes a multifaceted construct within the health sciences, encompassing a range of paradigms with distinct methodological and epistemological interpretations [1]. Within this landscape, the generation of synthetic data has evolved into an indispensable instrument for addressing critical challenges, including the validation of analytical methodologies, the rigorous assessment of experimental design properties, and the enforcement of patient privacy protections [2]. The advancement of artificial intelligence has precipitated the integration of generative adversarial networks (GANs) into synthetic data workflows, wherein models are trained to approximate and sample from the empirical distributions inherent in authentic datasets [3]. This has led to a rapid expansion of GAN-based applications spanning tabular records, photographic imaging, and radiological data modalities [4]. Long before the emergence of GAN-driven paradigms, statistical methodologies for synthetic data generation were established on a multi-decadal foundation. The literature delineates three archetypal classes of synthetic datasets: those constructed from real data, those generated independently of real data, and hybrid configurations that amalgamate both approaches [5]. The conceptual blueprint for substituting real data with synthetic surrogates was first articulated over three decades ago by Rubin, and it remains deeply rooted in statistical imputation theory [6]. Subsequent developments in statistical science yielded techniques tailored to each class; for instance, inverse-transform sampling and composition methods facilitate the creation of entirely artificial datasets absent any reliance on empirical observations [7]. When the objective is to evaluate a novel method under conditions of limited real-world sample sizes, synthetic data derived from genuine observations become pivotal. In such simulation studies, the verisimilitude between synthetic and real data is of paramount importance. This fidelity necessitates the accurate inference and reproduction of the complex data-generating process, particularly the joint distribution of multivariate datasets harboring mixed continuous and categorical variables a task that continues to pose substantial methodological challenges. A considerable corpus of statistical research has been devoted to surmounting these complexities across diverse real-world health data scenarios [8].

Also the computer science discipline has leveraged GAN-based generation to overcome obstacles related to small sample sizes, facilitate privacy-preserving data sharing, and reduce expenditure for example, through the synthesis of placebo group results [9]. Privacy preservation remains a cardinal motivation for deploying synthetic data in health contexts [10], spurring the development of numerous GAN architectures. Theoretically, strict privacy guarantees can be furnished through differential privacy frameworks, which ensure that the learning algorithm captures population-level information while disclosing nothing about any

individual. This principle has been operationalized in architectures such as PATE-GAN, where the discriminator aggregates outputs from an ensemble of teacher models trained on disjoint data partitions to produce a differentially private probability estimate [11]. Alternative strategies, though not grounded in formal theoretical proofs, offer practical privacy protections; an exemplar is ADS-GAN, which optimizes against a defined re-identification risk metric representing the likelihood of singling out an individual record given the entirety of the synthetic sample [12]. Benchmarking investigations have systematically evaluated the efficacy of various GAN-based generators across distinct tasks [13]. Notably, the conditional generative adversarial network tailored for tabular data (CTGAN) has emerged as a widely adopted solution in health research, owing to its capacity to model non-Gaussian marginal distributions and intricate inter-column dependencies [14]. Focusing specifically on tabular health data where each observation is represented as a row and each attribute as a column a fundamental question persists: does a GAN-based approach confer tangible advantages over classical statistical methods for synthetic data generation? To address this unresolved question, the present study first systematically maps the existing corpus of review studies on synthetic data generation methodologies. Subsequently, we conduct an empirical comparison concentrating on the generation of synthetic health tabular data within small-to-moderate sample size regimes, foregrounding practical implementation considerations. The classical statistical pipeline investigated combines the Mock Data Generation (Modgo) approach to simulate the feature matrix while preserving the empirical dependence structure, coupled with a parametric model-based strategy to simulate the outcome variable [15]. Modgo was selected based on its documented efficacy with tabular health data and its operational ease of implementation; competing statistical frameworks such as Faux, SimChef, SimMultiCorrData, and SimCorrMix frequently require the pre-specification of distributional parameters or known correlation matrices [16]. For the GAN-based paradigm, we employ the CTGAN architecture, which has demonstrated robust performance and widespread adoption in health-related studies [3]. It is important to underscore that the present comparison does not directly juxtapose Modgo with CTGAN as complete, end-to-end generation pipelines, because within the statistical framework the outcome variable is simulated through an independently specified model-based process. This investigation distinguishes itself from preceding reviews by delivering a cross-disciplinary, practically grounded evaluation that illuminates implementation choices pertinent to specific research goals. It thereby creates an avenue for the statistical community to adopt synthetic data generation techniques originating from computational sciences and provides a real-data-driven evaluation through detailed, application-oriented comparisons. The empirical assessment encompasses real-world datasets containing a mixture of categorical and continuous covariates, alongside continuous, binary, and time-to-event outcomes. The remainder of this manuscript is organized as follows. Section 2 presents a condensed review of existing methods, situating them within the evolving research landscape. Section 3 delineates the design of the empirical investigation. Section 4 reports the findings, first interrogating the similarity between synthetic and real data and thereafter addressing salient practical considerations. Section 5 concludes with a discussion of the respective strengths and limitations of statistical and GAN-based approaches and offers actionable recommendations for practitioners.

2. Evidence Map of Growing Popularity

To illuminate the trajectory of recent progress and the rapid expansion within synthetic data generation, a systematic search of the Scopus database was executed. The query incorporated the terms ‘data generation’, ‘data simulation’, ‘synthetic data generation’, ‘synthetic data simulation’, and ‘simulation study’, spanning January 2000 to October 2025, with non-English publications excluded. The analysis revealed a pronounced acceleration in publication volume after 2020 within the domains of ‘data generation’ and ‘synthetic data generation’ (depicted by the red and green trajectories in Figure 1(a)). In marked contrast, the trend for ‘data simulation study’ (pink trajectory) remained comparatively static over the quarter-century

period, indicating only a steady, modest increment in scholarly output within this subfield. A deeper examination of review studies was conducted, taking the year 2024 as an illustrative example using the search term 'synthetic data generation'; fourteen review studies were identified, yet none provided a practical performance comparison of the methods surveyed. To survey the methodological landscape specifically targeting tabular health data, an additional Scopus database search was performed using the terms 'synthetic data generation' AND 'tabular' OR 'microdata' OR 'micro' OR 'health tabular'. The top five predominant disciplines were visualized for the period 2023 to 2025 in Figure 1b. While the computer science domain consistently commands the majority of publications, a notable expansion was observed within the mathematics domain from 2023 to 2025. Remarkably, the proportional representation of studies within medicine exhibited a substantial increase in 2025 relative to 2023 and 2024, signaling a burgeoning interest in the application of advanced synthetic data techniques to health-related challenges. A manual curation of methods was subsequently undertaken from the identified review studies [8] and from the R CRAN repository (https://cran.r-project.org/), selecting only those approaches for which well-maintained application packages are available; these are catalogued in Table 1. Acknowledging the heterogeneous taxonomies adopted across different review studies, the compiled methods are not rigidly partitioned into classical statistical or GAN-based categories but are instead presented as a comprehensive list. Readers are directed to the original publications cited within Table 1 for detailed technical specifications of each listed method.

3. Empirical Study Design

3.1. Datasets

A total of six publicly available datasets, widely adopted in benchmarking studies, were selected to encompass a heterogeneous mixture of covariate types and distinct downstream analytical tasks. Specifically, two datasets were designated for each of the following outcome classes: linear regression, binary classification, and survival analysis. The characteristics of three representative datasets are summarized in Table 2; the remaining three are detailed in the Supplementary Material.

a) Vitamin D Data

The Vitamin D dataset was derived from the National Health and Nutrition Examination Survey (NHANES) [38]. Serum vitamin D concentration served as the continuous response variable, and a set of p=13 predictors was adopted based on prior investigations by Daraghmeh and Wang [39,40]. The complete dataset comprises n=10,195 observations.

b) Heart Disease Data

The heart disease dataset originated from a study by [17]. The binary target variable distinguishes between the presence and absence of heart disease. The predictor space includes a combination of continuous and categorical covariates, with p=13 and a sample size of n=1,025.

c) Veteran's Lung Cancer Data

The Veteran's Administrative Lung Cancer dataset consists of n=137 patients diagnosed with advanced inoperable lung cancer [18]. This right-censored survival dataset includes 7 predictor variables. The outcome of interest is survival time measured from study entry to the last follow-up, with patient death recorded as the event indicator.

3.2 Synthetic Data Generation Methods

The generation of synthetic data was performed using two distinct paradigms: a classical statistical pipeline and a generative adversarial network-based method employing CTGAN.

a) Classical Statistical Approach

The statistical pipeline was executed in two sequential stages: simulation of the covariate matrix, and simulation of the outcome variable.

(i) ***Simulation of the covariate matrix:*** The objective of this stage was to produce a synthetic covariate matrix X that faithfully replicates the dependence structure observed in the real data. Numerous statistical techniques exist for simulating mixed continuous-categorical covariates from empirical data [12]; the present study employs the recently developed Mock Data Generation (Modgo) methodology [19]. To evaluate the impact of feature standardization on simulation fidelity, the covariate matrix was generated both with and without a preliminary scaling step that transforms all columns of X to zero mean and unit variance.

(ii) Simulation of the outcome variable: Outcome variables were synthesized via a parametric model-based strategy [20]. For the continuous case, a linear regression model was first fit to the full real dataset to obtain the regression coefficient estimates. The previously generated synthetic covariate matrix X was then combined with these coefficients and a standard normally distributed random error term to simulate the response. For the binary outcome, a logistic regression model was employed analogously, and for the censored survival outcome, a Cox proportional hazards model was adopted [4]. While a broader array of generative models could be envisioned, such an extension lies beyond the scope of this work.

b) CTGAN Approach

The CTGAN architecture was deployed with the number of training epochs set to 1000, a value generally considered sufficiently large to ensure convergence and robust performance [16]. In contrast to the statistical pipeline, the CTGAN framework does not necessitate the intermediate step of estimating model-based coefficients; it learns the entire joint data distribution directly from the real observations.

3.3 Method Performance Comparison

A distinctive strength of the present investigation, relative to existing comparative studies, is the simultaneous integration of three disparate health outcome types with a systematic assessment of synthetic data quality coupled with practical implementation considerations. For each combination of dataset and generation method, 100 independent synthetic datasets were produced and evaluated.

The comparison criteria were organized into two principal categories.

(a) Data Similarity

Quantitative metrics: For every continuous variable, the sample mean, variance, skewness, and kurtosis were computed as fundamental moment-based similarity indicators [12]. A one-sample t-test was subsequently applied to assess whether the real-data value of each moment differed significantly from the distribution of 100 simulated estimates. Distributional congruence for continuous variables was further scrutinized using a suite of metrics aligned with recent benchmarking guidelines [8]: the inverse of the Kullback–Leibler divergence, the Kolmogorov–Smirnov statistic, the Jensen–Shannon distance, and the Wasserstein distance. The concordance of categorical variable distributions was evaluated via the Chi-square test, and pairwise associations were quantified using Pearson correlation coefficients.

Qualitative metrics: Visual diagnostics complemented the quantitative comparisons. Histograms were employed to compare univariate continuous distributions, bar plots were generated for categorical frequencies, and heatmaps illustrated the structure of inter-variable correlations. Additionally, bivariate relationships were examined through two-dimensional density plots.

(b) Prediction Performance Measurements

The utility of synthetic data for downstream modeling was evaluated by fitting prediction models on the synthetic datasets and measuring their accuracy when applied to the original real data [8]. For the linear regression setting, prediction accuracy was quantified using the mean squared error (MSE) [21]; lower MSE values indicate superior performance. For binary classification, balanced accuracy served as the metric, while the concordance index (C-index) was employed for the Cox survival model [22]; in both cases, larger values correspond to better discriminative performance. The alignment of estimated model coefficients derived from synthetic data with those obtained from the real data was assessed using the interval-overlap metric [22].

4. Comparison Results

4.1. Similarity of Synthetic Data to Real Data

a) Examination of Descriptive Statistics

The initial analysis focused on the congruence of fundamental descriptive statistics between synthetic and real datasets. As anticipated, the sample means were highly concordant across both generation methods; however, higher-order moments particularly skewness and kurtosis exhibited noticeable discrepancies.

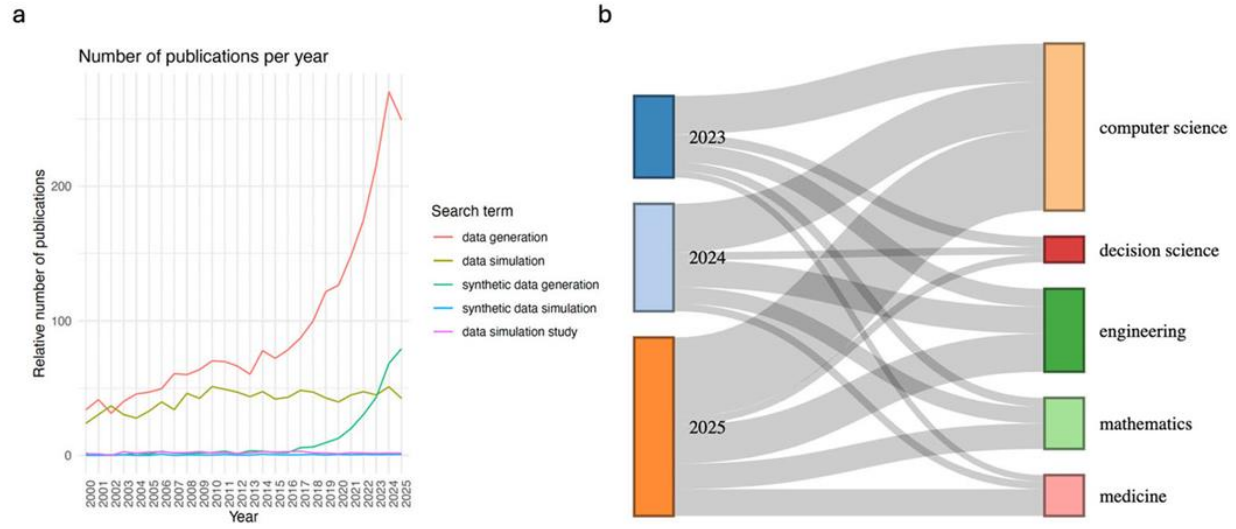

Figure 1. Review of the literature. (a) Relative number of publications per year, calculated as the number of publications divided by the total number of publications in each year and scaled by 10k. The total number of publications per year is defined by restricting the year range and lan guage to be 'English' in the search. (b) Areas of publications in 2023, 2024 and 2025 (to October 20th 2025).

Variances demonstrated dataset-dependent behavior: the synthetic and real variances were closely aligned for the Heart dataset, whereas for the Vitamin D and Veteran datasets, notable divergences emerged. Formal evaluation via one-sample t-tests revealed no statistically significant differences (at the 5% significance level) between the real-data value and the distribution of synthetic estimates for nearly all variable means. For categorical variables, initial visual inspection using bar plots indicated no overt deviations between the real and synthetic distributions. Corroborating these visual assessments, Chi-square tests of equality of category proportions similarly failed to detect statistically significant differences between the real data and the synthetic data produced by either method. No pronounced data-dependent performance patterns were observed; consistent findings were replicated across the three additional datasets analyzed in the Supplementary Material.

(b) Examination of Distributions

A detailed comparison of univariate and bivariate distributions was conducted to evaluate the fidelity of the statistical and CTGAN approaches. Using the four quantitative distributional similarity metrics defined in Section 3.3(i), both methods were found capable of generating variables with distributions closely approximating those of the real data. Bivariate dependence structures were further examined using two-dimensional density visualizations, which illustrated the relationships between pairs of variables across three configurations: continuous–continuous, continuous–categorical, and continuous–binary outcome. The synthetic datasets produced by both approaches exhibited density contours that closely mirrored those of the real data, demonstrating an ability to capture the joint behaviour of mixed-type covariates as well as the association between covariates and the outcome variable. For instance, the tendency of a continuous variable to cluster around a central value conditional on a categorical variable was consistently reproduced in the synthetic data, reflecting high fidelity. However, a subtle difference was observed: the CTGAN-generated densities occasionally displayed a more discretized or fragmented appearance, with increased nucleation, a phenomenon particularly pronounced in the Veteran dataset. These findings indicate that while both paradigms effectively learn the global dependence patterns, the statistical method yields smoother joint representations for certain data structures.

(c) Examination of Correlations

Accurate reproduction of the multivariate correlation structure is a fundamental requirement in synthetic data generation. The performance of the two approaches was therefore compared with respect to their ability to preserve pairwise correlations among variables. Across all

datasets examined, the correlation matrices derived from the statistical synthetic data exhibited a closer correspondence to those of the real data than did those from the CTGAN synthetic data. A representative example is the correlation between the variables num_karno and fac_prior in the Veteran dataset, which was more faithfully recovered by the statistical approach, whereas the CTGAN method produced a notably inflated correlation estimate. This outcome is theoretically anticipated: the statistical pipeline explicitly estimates the empirical correlation structure from the real data and applies analytical formulae designed to replicate it in the generated covariate matrix. In contrast, CTGAN learns the joint data distribution implicitly through the adversarial interplay of a generator and a discriminator, with model parameters updated via back-propagation based on the discriminator's output probability. While this process can capture complex dependencies, it does not directly enforce exact replication of pairwise linear correlations, leading to partial preservation. Despite this relative divergence, the overall correlation patterns produced by CTGAN remained broadly consistent with the real data, with deviations generally limited in magnitude.

4.2 Practical Considerations

4.2.1 Data Pre-processing

Appropriate data preparation prior to synthetic data generation is critical to the success of the employed methodology. This investigation demonstrates that the statistical approach exhibits pronounced sensitivity to variable scaling, whereas the CTGAN approach proves substantially more robust. Focusing on the VitD dataset and the variables Calcium, Sodium, and the outcome vitD, it was observed that, in the absence of prior scaling, the statistical method failed to reproduce the real data distributions accurately. Most notably, for the outcome variable vitD, the unscaled statistical pipeline generated values spanning an entirely incorrect range. After applying a standardization step (centering and scaling to unit variance), the statistical method's performance improved markedly, yielding synthetic distributions in close alignment with the real data. By contrast, the distributions of variables synthesized by CTGAN exhibited high concordance with the real data irrespective of whether scaling was applied. This outcome underscores the CTGAN method's lower vulnerability to the scale of input features, which may confer practical advantages when the optimal preprocessing protocol is uncertain.

4.2.2 Downstream Analysis

Synthetic data are frequently employed in statistical model development and predictive analytics. To evaluate utility in such contexts, prediction models were trained on the synthetic datasets and their predictive accuracy was assessed on the held-out real data. A slight performance advantage was observed for the statistical approach in binary classification and survival analysis tasks, whereas the CTGAN approach offered a modest benefit for linear regression. Specifically, using the Heart dataset, the balanced accuracy achieved with models trained on statistical synthetic data was 0.85 (SD = 0.01), compared with 0.78 (SD = 0.01) for CTGAN. For the Veteran survival data, the concordance index was 0.71 (SD = 0.004) for the statistical method and 0.65 (SD = 0.042) for CTGAN. Conversely, on the VitD regression task, the statistical synthetic data yielded a mean squared error of 2.82 (SD = 2.59), whereas CTGAN produced a notably lower value of 1.02 (SD = 0.06). Examination of the interval-overlap metric for estimated model coefficients revealed that, for the majority of variables in the VitD and Heart datasets, the statistical approach provided superior coefficient recovery, whereas for the Veteran dataset the CTGAN method performed better. These differences can be rationalized by the contrasting generation mechanisms: the statistical paradigm explicitly incorporates the parametric relationship between covariates and the outcome, while CTGAN learns this association indirectly as part of the full joint distribution. Despite their distinct underpinnings, the two approaches demonstrated broadly comparable predictive performance, a finding that consistently replicated across the additional three datasets analyzed.

Table 3: *Prediction performances (based on 100 repeats).*

Dataset	Prediction model	Metric	Statistical approach (mean ± SD)	CTGAN (mean ± SD)
VitD	Linear regression	MSE	2.82 (2.59)	1.02 (0.06)
Heart	Logistic regression	Balanced accuracy	0.85 (0.01)	0.78 (0.01)
Veteran	Cox model	C-index	0.71 (0.004)	0.65 (0.042)

*Note: MSE = Mean Squared Error (lower values indicate better performance); Balanced accuracy and C-index are reported with higher values reflecting superior discrimination. Standard deviations are shown in parentheses, computed over 100 independently generated synthetic datasets.

5. Discussion

The accelerating convergence of generative artificial intelligence and statistical science has engendered a paradigm shift in the conceptualization and execution of simulation studies within the health sciences. The present investigation was conceived in direct response to this transdisciplinary momentum, aiming to furnish a rigorous, empirically grounded evaluation of whether adversarial learning frameworks can serve as viable substitutes for or complements to classical parametric simulation pipelines. By juxtaposing a state-of-the-art conditional generative adversarial network (CTGAN) against a modern statistical method (Modgo) with model-based outcome imputation, across six real-world tabular datasets spanning linear, logistic, and Cox regression contexts, we have generated a nuanced body of evidence [23]. The results demonstrate that both methodological families are capable of recovering the salient univariate distributional features and complex multivariate dependence architectures inherent in mixed continuous-categorical health records, and that they yield downstream predictive models of largely comparable discriminative and calibration performance. These empirical observations collectively dismantle the presumption that deep generative models are inherently unsuited to the stringent fidelity requirements of statistical simulation, and instead position them as credible instruments within the applied researcher's methodological toolkit.

Notwithstanding this convergence in performance, our findings illuminate critical dissimilitude in the generative mechanisms, failure modes, and operational prerequisites of the two paradigms insights that carry profound implications for principled method selection. The classical statistical pipeline, which decouples covariate synthesis from outcome generation, relies on the explicit estimation and analytical propagation of a dependence structure. In Modgo, the empirical correlation matrix is computed and subsequently enforced through a copula-based or moment-matching algorithm, ensuring that pairwise linear associations are reproduced with high numerical fidelity [24]. This property was empirically corroborated: the correlation heatmaps derived from statistical synthetic data exhibited greater congruence with the real data than those produced by CTGAN, most notably in datasets such as Veteran where certain inter-variable relationships were markedly inflated by the adversarial generator. The statistical approach thus retains an unequivocal advantage in scenarios where the veridical reproduction of a pre-specified correlation network is of paramount scientific import for instance, when the covariance structure encapsulates prior biological knowledge, is derived from a validated structural equation model, or when the simulation serves as a sensitivity analysis for a mediation or instrumental variable design. Furthermore, the explicit separation of feature and outcome simulation permits the analyst to impose counterfactual perturbations on the parametric outcome model, interrogate the impact of model misspecification in a controlled fashion, and quantify the propagation of uncertainty through identifiable analytical pathways [25]. These capabilities, deeply rooted in the inferential logic of classical statistics, remain challenging to replicate within the black-box optimization regimen of generative adversarial networks. Conversely, the CTGAN architecture eschews explicit correlation targeting in favor of an implicit, distribution-wide adversarial learning objective. The generator is trained to transform samples from a latent prior into synthetic vectors that a discriminator cannot reliably distinguish

from real data. This adversarial loss, formulated as a minimax game, induces the generator to approximate the full joint probability density function, including non-linear dependencies, higher-order interactions, and complex conditional heterogeneities that lie beyond the reach of linear correlation coefficients. Our empirical observation that CTGAN occasionally produced slightly discretized or nucleated density contours most pronounced in the small-sample Veteran data can be theoretically rationalized as a manifestation of mode collapse mitigation strategies operating under the constraints of limited data. While a pure mode-seeking generator might collapse to a few high-density regions, CTGAN's use of conditional vectors and training-by-sampling techniques encourages the exploration of minor modes, albeit at the cost of introducing granular artifacts [26]. This discrete-like behavior, rather than being a defect, may in fact constitute an adaptive response to genuine subpopulation clustering, making CTGAN the method of choice when variables exhibit pronounced multimodality, heavy tails, or zero-inflation that would severely strain standard parametric families. The method's non-parametric flexibility thus positions it as a powerful option for exploratory simulation contexts where the underlying data-generating process is poorly characterized or suspected to deviate from classical distributional assumptions [17].

A decisive practical differentiator illuminated by our study concerns the sensitivity to data preprocessing. The statistical pipeline's reliance on moment-based transformations renders it acutely vulnerable to scale heterogeneity among variables. When features span orders of magnitude, unscaled data cause the correlation estimation and subsequent matrix decomposition steps to be dominated by high-variance columns, resulting in catastrophic distortion of the synthetic distributions as starkly illustrated by the complete failure to reproduce the outcome range for vitamin D without standardization [21]. This implies that the statistical approach demands a vigilant data inspection phase: analysts must verify the empirical range and distributional shape of every variable, apply appropriate variance-stabilizing transformations, and consider robust correlation estimation methods in the presence of outliers. Moreover, the downstream outcome simulation requires the specification of a parametric model; a misspecified link function, the omission of a non-linear term, or an incorrect distributional assumption for the error term will propagate systematic bias into the entire synthetic dataset, compromising its utility for both inference and prediction. In contrast, CTGAN demonstrated remarkable robustness to the scale of raw input features, generating distributions of consistently high fidelity irrespective of standardization. This insensitivity stems from the neural network's inherent capacity to learn scaling transformations through gradient-based optimization, as well as from the use of batch normalization and data-type-specific nonlinearities within its architecture. This characteristic substantially reduces the pre-analytical burden and the risk of human error, making CTGAN an appealing solution in high-throughput data environments or for analysts with limited expertise in statistical simulation [23]. This procedural simplicity is offset by the empirical and computationally intensive tuning demands of the adversarial training process. The number of training epochs, in the absence of a formal convergence guarantee analogous to the monotonic likelihood improvement in expectation–maximization or Markov chain Monte Carlo diagnostics, must be determined through a heuristic grid search monitoring the trajectory of synthetic data quality metrics. In this work, following established precedents, we evaluated epoch numbers ranging from 100 to 2000, ultimately selecting 1000 as a pragmatic equilibrium between generative fidelity and computational expenditure. Yet this choice is inherently dataset-dependent; large, high-dimensional tables may require substantially longer training to capture rare combinatorial patterns, while small datasets risk overfitting with excessive epochs, whereby the generator memorizes and regurgitates training instances a phenomenon closely tied to privacy vulnerabilities [27]. The absence of a universally applicable, theory-grounded stopping criterion for GAN training represents a salient lacuna in the current methodological literature and a non-trivial impediment to the fully automated deployment of these tools in regulatory-grade simulation workflows. Future research should

prioritize the development of diagnostic frameworks, perhaps based on the stability of sliced Wasserstein distances between consecutive synthetic batches or on the convergence of discriminator gradients, to provide objective and interpretable termination rules. It is imperative to contextualize the predictive performance results within the broader landscape of synthetic data utility. The marginal superiority of the statistical approach in binary classification (balanced accuracy 0.85 vs. 0.78) and survival analysis (C-index 0.71 vs. 0.65) may be attributable to the explicit parametric incorporation of the outcome-covariate relationship during data generation, which imposes an inductive bias well-aligned with the subsequent model-fitting exercise. CTGAN, by learning the joint distribution without privileging any variable as the outcome, must additionally expend model capacity on recovering dependencies among covariates that are ancillary to the prediction task, potentially diluting its focus on the discriminative boundary [28]. Conversely, for continuous outcomes in linear regression, CTGAN's lower mean squared error (1.02 vs. 2.82) could reflect its ability to capture subtle non-linearities or heteroscedasticity that the rigid linear model-based simulation fails to encode, thereby providing a richer training environment that improves generalization when the real data themselves contain such complexities. The interval-overlap metric further nuanced this picture: the statistical method better recovered coefficients for most variables in the VitD and Heart datasets, while CTGAN excelled in the Veteran dataset an observation that may be linked to the small sample size and the corresponding difficulty of accurately estimating parametric model coefficients, a scenario where the implicit regularization of adversarial training might confer an advantage. These dataset-dependent variations underscore a cardinal principle: the optimal synthetic data generation strategy is contingent upon the specific inferential or predictive objective, the sample size, and the latent complexity of the data-generating mechanism [29]. A deliberate boundary condition of our study is the exclusion of privacy preservation from the evaluation framework. The quantification of privacy risk in synthetic data is a domain fraught with theoretical and practical challenges. There exists no universally endorsed metric; instead, the field is fragmented into two broad paradigms. The first assesses robustness against canonical adversarial attacks, most notably membership inference attacks, wherein an adversary with access to the synthetic data and a query interface seeks to determine whether a particular record was included in the training set. This evaluation typically involves the computation of a distance or likelihood ratio threshold, the calibration of which is highly sensitive to the attack model and the dimensionality of the data. The second paradigm quantifies re-identification risk through population-level similarity metrics, such as ℓ-diversity, k-map, or the propensity-score-based measures integrated within specialized frameworks like ADS-GAN and synthcity. These metrics estimate the probability that a real individual's sensitive attributes can be uniquely linked to a synthetic record, given an adversary's auxiliary knowledge. While CTGAN itself lacks intrinsic privacy guarantees, its outputs can be post-processed or the architecture can be augmented with differential privacy mechanisms such as PATE-GAN to provide formal bounds. The omission of privacy considerations from the present work should not be construed as a dismissal of their importance; rather, it acknowledges that privacy evaluation constitutes a complex, multi-dimensional research question that merits a dedicated, standalone investigation with its own set of design decisions, threat models, and performance trade-offs. The integration of privacy-enhancing technologies with the quality-utility benchmarks established herein represents a critical next step toward the deployment of synthetic health data in sensitive, regulated environments. The landscape of both generative AI-based and classical statistical simulation methods is vast, dynamic, and continually enriched by novel contributions. Our study has deliberately focused on a single, well-regarded representative from each domain; it has not sought to exhaustively benchmark the multitude of available algorithms, such as the variants of variational autoencoders, diffusion models, normalizing flows, or alternative statistical copula and sequential imputation techniques. Each of these methods carries distinct assumptions, strengths, and limitations

that may interact with dataset characteristics in unpredictable ways. Comprehensive intra-domain comparison studies exist and serve as valuable resources for methodologists. The contribution of the present work lies instead in its cross-paradigm, practically oriented empirical framework that emphasizes real-world tabular health data with mixed outcomes, small to moderate sample sizes, and actionable guidance for the end-user. By articulating the circumstances under which each family of methods is likely to excel or falter informed by both theoretical principles and empirical evidence we equip the research community with a decision-making heuristic that can be refined as new methodologies emerge [30].

Looking forward, several avenues demand concerted investigation. First, the development of hybrid architectures that synergistically combine the structural interpretability and parametric control of statistical models with the distributional flexibility of deep generative networks holds immense promise. For instance, a copula-based generator could be used to model marginal distributions and pairwise dependencies within a probabilistic graphical model framework, while a residual adversarial network could learn and inject the remaining higher-order interactions that elude the copula. Second, the establishment of a standardized, multi-faceted evaluation protocol for synthetic health data encompassing fidelity, utility, privacy, fairness, and diversity would greatly facilitate reproducibility and method comparison across studies. Third, the theoretical properties of CTGAN-like models under finite-sample and high-dimensional asymptotic regimes require deeper formalization, particularly regarding the trade-off between data memorization and generalization. Finally, empirical investigations in causal inference settings, where synthetic data must preserve not only associational but also interventional distributions, will be essential to assess the fitness of these tools for the next generation of health data science. As generative AI continues its trajectory of rapid advancement, the symbiotic integration of rigorous statistical principles will be indispensable to ensure that synthetic data not only mimic reality but do so in a manner that is safe, scientifically valid, and ethically sound.

Conflict of interest

The authors declared no conflict of interest.

Author Contribution

All authors reviewed the results and approved the final version of the manuscript. They are also accountable for the study's integrity.

References

1. Parise O, et al. CTGAN-driven synthetic data generation: a multidisciplinary, expert-guided approach (TIMA). Comput Methods Programs Biomed. 2025;259:108523. doi:10.1016/j.cmpb.2024.108523
2. Nowok B, et al. synthpop: generating synthetic versions of sensitive microdata for statistical disclosure control. CRAN: Contributed Packages. 2025. doi:10.32614/CRAN.package.synthpop
3. DeBruine LF. Simulation for factorial designs. CRAN: Contributed Packages. 2025. doi:10.5281/zenodo.2669586, R package version 1.2.3.
4. Friedrich S, Friede T. On the role of benchmarking data sets and simulations in method comparison studies. Biom J. 2024;66(1):2200212. doi:10.1002/bimj.202200212
5. Duncan J, Tang T, Elliott CF, et al. High-quality data science simulations in R. J Open Source Softw. 2024;9(95):6156. doi:10.21105/joss.06156
6. Bauer A, Trapp S, Stenger M, et al. Comprehensive exploration of synthetic data generation: a survey. arXiv preprint arXiv:2401.02524. 2024 Jan 4.
7. Qian Z, Davis R, Van Der Schaar M. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. Adv Neural Inf Process Syst. 2023;36:3173–3188.
8. Kang HY, Batbaatar E, Choi DW, et al. Synthetic tabular data based on generative adversarial networks in health care: generation and validation using the divide-and-conquer strategy. JMIR Med Inform. 2023;11:e47859. doi:10.2196/47859
9. Sun C, van Soest J, Dumontier M. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. J Biomed Inform.

2023;143:104404. doi:10.1016/j.jbi.2023.104404
10. Koliopanos G, Ojeda F, Ziegler A. A simple-to-use R package for mimicking study data by simulations. Methods Inf Med. 2023;62(03/04):119–129. doi:10.1055/a-2048-7692
11. Norcliffe A, Cebere B, Imrie F, et al. Survivalgan: generating time-to-event data for survival analysis. In: International Conference on Artificial Intelligence and Statistics; 2023 Apr 11. p. 10279–10304. PMLR.
12. Watson DS, Blesch K, Kapar J, et al. Adversarial random forests for density estimation and generative modeling. In: International Conference on Artificial Intelligence and Statistics; 2023 Apr 11. p. 5357–5375. PMLR.
13. Wang TY, Wang HW, Jiang MY. Prevalence of vitamin D deficiency and associated risk of all-cause and cause-specific mortality among middle-aged and older adults in the United States. Front Nutr. 2023;10:1163737. doi:10.3389/fnut.2023.1163737
14. Ghosheh G, Li J, Zhu T. A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources. arXiv preprint arXiv:2203.07018. 2022 Mar 14.
15. Hernandez M, Epelde G, Alberdi A, et al. Synthetic data generation for tabular health records: a systematic review. Neurocomputing. 2022;493:28–45. doi:10.1016/j.neucom.2022.04.053
16. Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs. Math. 2022;10(15):2733. doi:10.3390/math10152733
17. Kraft R, Martínez Madrid N, Seepold R. Generative adversarial networks: project relevant overview. Hochschule Reutlingen. 2022:23–25.
18. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc J. 2022;9(2):190–193. doi:10.7861/fhj.2022-0013
19. Zhang Y, Wong G, Mann G, et al. SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data. Gigascience. 2022;11:giac071. doi:10.1093/gigascience/giac071
20. Cai S, Zhao L, Ban Y, et al. GAN-based image-to-friction generation for tactile simulation of fabric material. Comput Graph. 2022;102:460–473. doi:10.1016/j.cag.2021.09.007
21. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med. 2022;1(1):e000167. doi:10.1136/bmjmed-2022-000167
22. Boulesteix AL, Groenwold RH, Abrahamowicz M, et al. Introduction to statistical simulations in health research. BMJ Open. 2020;10(12):e039921. doi:10.1136/bmjopen-2020-039921
23. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–2102. doi:10.1002/sim.8086
24. Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional gan. Adv Neural Inf Process Syst. 2019;32.
25. Yoon J, Drumright LN, Van Der Schaar M. Anonymization through data synthesis using generative adversarial networks (ads-gan). IEEE J Biomed Health Inform. 2020;24:2378–2388. doi:10.1109/JBHI.2020.2980262
26. Baowaly MK, Liu CL, Chen KT. Realistic data synthesis using enhanced generative adversarial networks. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE); 2019 Jun 3. p. 289–292. IEEE.
27. Snoke J, Raab GM, Nowok B, et al. General and specific utility measures for synthetic data. J R Stat Soc A: Stat Soc. 2018;181(3):663–688. doi:10.1111/rssa.12358
28. Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction. Math Intell. 2005;27(2):83–85. doi:10.1007/BF02985802
29. Rubinstein RY, Kroese DP. Simulation and the Monte Carlo method. Hoboken (NJ): John Wiley & Sons; 2016 Oct 20.

30. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Stat Med. 2013;32(23):4118–4134. doi:10.1002/sim.5823