JAMIA Open, 2025, **8(6)**, ooaf137
https://doi.org/10.1093/jamiaopen/ooaf137
Research and Applications

Research and Applications

Synthetic data generation of health and demographic surveillance systems data: a case study in a low- and middle-income country

Dorcas G. Mwigereri, MSc[1,*], Nigel T. Kamotho, BSc[2], Akbar K. Waljee, MD, MSc[1,3,4], Ryan T. Rego, PhD[3,5], Eileen M. Weinheimer-Haus, PhD[3,4], Farhana Alarakhiya, BSc[2], Anthony K. Ngugi, PhD[1], W. Nicholson Price II, PhD[4,6], Ji Zhu, PhD[7], Stephen Peter Wong, BSc[2], Geoffrey H. Siwo, PhD[3,4]

[1]Department of Population Health, The Aga Khan University, 30270–00100 Nairobi, Kenya, [2]Data Innovation Office, Aga Khan Univeristy, 30270–00100 Nairobi, Kenya, [3]Center for Global Health Equity, Michigan Medicine, University of Michigan at Ann Arbor, Ann Arbor, MI 48109, United States, [4]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI 48109, United States, [5]Maximum Impact Incubator, Clinton Health Access Initiative, Boston, MA 02127, United States, [6]University of Michigan Law School, Ann Arbor, MI 48109, United States, [7]Department of Statistics, University of Michigan, Ann Arbor, MI 48109, United States

*Corresponding author: Dorcas G. Mwigereri, MSc, Department of Population Health, The Aga Khan University, 30270–00100 Nairobi, Kenya (dorcas.mwigereri@aku.edu)

D.G. Mwigereri and N.T. Kamotho are co-first authors of this work.

S.P. Wong and G.H. Siwo are co-senior authors of this work.

Abstract

Objective: To evaluate effectiveness of open-source generative models in producing high-quality tabular synthetic data using a Health and Demographic Surveillance System (HDSS) dataset from rural Kenya, as a proof of concept in a low- and middle-income (LMIC) setting.

Materials and Methods: Three open-source models (CTGAN, TableGAN, and CopulaGAN) were used to generate synthetic data from the Kaloleni/Rabai HDSS dataset. To assess the quality of the synthetic datasets generated by each model, we performed fidelity, utility, and privacy tests.

Results: CTGAN outperformed the other models, producing synthetic data that closely mirrored the statistical properties of the real dataset while preserving privacy. Both CopulaGAN and TableGAN performed poorly, with TableGAN completely failing to generate realistic synthetic data. For the utility tests, Random Forest models trained on CTGAN-generated synthetic data achieved comparable performance to models trained on real data (accuracy: 72.4% vs 72.0%, $P = .38$; F1 score: 71.4% vs 68.3%, $P = .22$), indicating no statistically significant loss in predictive utility. The CTGAN model also yielded higher precision and recall than CopulaGAN, suggesting that the synthetic data generated by CTGAN better preserved the underlying structure of the real data.

Discussion: CTGAN demonstrated superior performance in generating high-quality synthetic tabular HDSS data. CopulaGAN and TableGAN produced lower quality data, though these results may not generalize to other datasets.

Conclusion: Synthetic data generation of tabular data using HDSS data, particularly via CTGAN, may enhance the accessibility of datasets in LMICs by creating synthetic datasets that preserve the characteristics and statistical properties of the original data, while upholding privacy and confidentiality.

Lay Summary

This study explored the effectiveness of open-source generative models in generating realistic, privacy-preserving synthetic datasets of a Health and Demographic Surveillance System (HDSS) dataset obtained from rural Kenya, as a proof of concept for low- and middle-income countries (LMICs). Using 3 open-source models—CTGAN, TableGAN, and CopulaGAN—synthetic data were created from the real data while maintaining important characteristics like missing values and outliers. The quality of the synthetic data was evaluated to determine whether it maintained the statistical properties present in the real data and provided privacy protection. Among the models tested, CTGAN performed the best, producing synthetic data that closely resembled the real data while preserving data privacy. In contrast, CopulaGAN and TableGAN were less successful, with TableGAN completely failing to generate realistic synthetic data. This study shows that CTGAN can be a valuable tool for creating synthetic data in LMICs, making real datasets more accessible and preserving privacy.

Key words: synthetic data; generative adversarial networks; health and demographic surveillance systems; HDSS; AI; machine learning.

Received: June 23, 2025; **Revised:** August 11, 2025; **Accepted:** October 6, 2025

Introduction

Health and Demographic Surveillance Systems (HDSS) provide vital data on population health trends, disease burdens, and the effectiveness of public health interventions.[1,2] In low- and middle-income countries (LMICs), HDSS data are among the most timely and detailed sources of population health and demographic data regularly generated, contributing to a growing repository of longitudinal data that form a valuable resource for research, policy-making, and public health interventions.[3] Despite their value, broad utilization of HDSS datasets remains difficult due to regulatory concerns around privacy and confidentiality.[4,5] Varying data protection laws across countries further limit data sharing, making it difficult to harmonize multimodal datasets and apply advanced data science tools, such as machine learning (ML) and artificial intelligence (AI), to generate actionable insights that can inform data-driven health policies and optimize resource allocation.[6–8]

Synthetic data generation offers a potential strategy to enhance accessibility of datasets by creating synthetic datasets that preserve the characteristics and statistical properties of the original (ie, "real") data, while upholding privacy and confidentiality.[9] Synthetic data generation techniques can be broadly categorized as physical models and statistical models.[10,11] The physical models simulate the underlying processes that generate real data based on known scientific or domain-specific laws and as such they require explicit parameterization and often incorporate domain knowledge.[12] While physical models provide interpretability and can reflect causal mechanisms, they can be costly, time-consuming, and less generalizable for complex real-world data. In contrast, statistical models for synthetic data generation, relying on observed probabilistic distributions, offer simplicity.[13] This can be achieved by using either hand-coded techniques based on expert knowledge or inferred from the data using models such as flow models,[14] variational autoencoders (VAEs),[15] diffusion models,[16] and generative adversarial networks (GANs).[17]

While synthetic data generation has gained traction in various healthcare domains, its application within global health remains relatively underexplored.[18] Oncology, neurology, and cardiology were identified as the most frequently addressed areas with limited emphasis on global health contexts according to a scoping review by Rujas et al.[18] Additionally, from their review, it was noted that most of the studies focus on generation of unstructured data such as images with less focus on other types of data such as texts, videos, and tabular data. While GANs have been widely used for synthetic image data generation, tabular GANs models have been adopted to extend the use of GANs for synthetic data generation given mixed data (both numerical and categorical data). The tabular GANs include TableGAN,[19] CTAB-GAN,[20] medGAN,[21] ITS-GAN,[19] CopulaGAN,[22] and CTGAN.[23] Xu et al[23] compared the performance of MedGan, VeeGan, and TableGAN with Bayesian networks for synthetic data generation. They found that these GAN techniques were less effective than Bayesian networks due to challenges in modeling tabular data and issues such as vanishing gradients. To address these problems, they proposed CTGAN, which uses a conditional generator and mode-specific normalization to handle class imbalance and improve performance, showing superior results compared to the other methods.

As proof of concept, this study evaluated the effectiveness of 3 open-source algorithms, CTGAN, TableGAN, and CopulaGAN, to generate synthetic data on HDSS data from the Kaloleni-Rabai Community HDSS (KRHDSS) in Kenya. Fidelity, utility, and privacy tests were performed to determine whether the synthetically generated HDSS data preserved important statistical properties presented in the real HDSS data and provided privacy protections.

Materials and methods

Dataset description

The KRHDSS is a comprehensive population-based health and demographic surveillance system established in 2017 by the Aga Khan University in collaboration with local health management in Kenya's Kaloleni and Rabai sub-counties.[24] This HDSS includes cross-sectional and longitudinal data on demographics, health events, and social determinants of health, with unique individual identifiers enabling longitudinal tracking and linkage for research purposes. Since its inception, 10 rounds of data collection have occurred. In this proof-of-concept study, we generated synthetic data using the latest rounds of data collection—round 6 (R6, July–December 2019) and round 8 (R8, July–December 2020)—which included 92 663 and 92 805 records, respectively. Data from multiple recorded tables were merged into a single table that included all member details and household information. The final dataset included 31 variables, all of which were categorical (Table S1). Missing values and outliers were retained for synthetic data generation to evaluate model performance using real-world HDSS data.

Synthetic data generation

Although various GAN methods exist for generating synthetic data, we focused on 3 open-source models—CTGAN,[23] TableGAN,[25] and CopulaGAN[26,27]—due to their capability to manage complex tabular data with intricate dependencies. The CTGAN[23] model uses a conditional generator to focus on the joint distribution of features, along with a generator loss module to penalize the generator and a training-by-sampling method to compare distributions. TableGAN,[25] in contrast, has a 3-part architecture comprising a discriminator to differentiate between real and synthetic data, a generator to obfuscate the discriminator, and a classifier neural network to enhance data integrity. Lastly, CopulaGAN[27] uses Gaussian Copulas to apply cumulative distribution function (CDF) transformations to capture the dependence structure between variables in a dataset while allowing for flexibility in modeling different marginal distributions for each variable.

Performance evaluation

Fidelity, utility, and privacy tests were used to evaluate the ability of the 3 algorithms to generate high-quality synthetic data as described below.

Fidelity

Fidelity tests evaluate whether the synthetic data accurately preserve key properties, patterns, and relationships that are present in the real data. Univariate distributions between the real and synthetic data were assessed by comparing the proportions of each individual feature. To evaluate the preservation of relationships between variables in synthetic data, we analyzed bivariate distributions through joint frequency distributions. Contingency tables were created for pairs of categorical variables, displaying the frequency of each combination. A heatmap was then employed to assess the consistency of relationships, providing a clear evaluation of how well the synthetic data preserved the relationships found in the real data.

Associations between variables were assessed using normalized mutual information (uncertainty coefficient) to evaluate how well the synthetic data preserves the associations between variables in the real data. To compare the distribution of associations between real and synthetic data, Kolmogorov-Smirnov (K-S) tests were performed to determine if these distributions were significantly different. The Wilcoxon rank-sum test was used to determine if the differences in the magnitude of associations in the real versus synthetic data were statistically significant.

Machine learning utility

We assessed the utility of synthetic data for machine learning models, ie, whether machine learning models trained on synthetic data and evaluated on real data perform similarly as those trained from real data. Specifically, we used the task of predicting whether an individual had access to a functional latrine ("latuse_r" with classes Yes/No) based on the other variables. Three instances of the Random Forest (RF) classifier were created, each trained on a different dataset: real data, synthetic data generated by CTGAN, and synthetic data generated by Copula using R6. All models were then evaluated on real data from R8 as illustrated in Figure 1.

Categorical features were one-hot encoded via a ColumnTransformer to ensure compatibility with the RF model. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC metrics.

Privacy

To assess privacy risk, we conducted an attribute inference attack using the Anonymeter framework[28] to determine whether synthetic data can be utilized by an attacker to accurately infer unknown attributes of individuals based on known information. This approach requires 3 datasets: a training set used to generate synthetic data, a control set held out from training for benchmarking, and the resulting synthetic dataset. The attack involves attempting to infer unknown attributes (secret variables) of individuals in the synthetic dataset, using a set of known attributes (auxiliary variables) available to the attacker.

Given that the KRHDSS training dataset does not contain personally identifiable information (PII), we selected the variable "hivstatus"—based on the survey question "Do you know your HIV status?"—as a proxy for a culturally sensitive attribute. While not inherently identifying, this variable carries contextual cultural sensitivity and has broad coverage across the population, making it a suitable stand-in for assessing potential risks of HDSS synthetic data.

We modeled 2 attacker scenarios. The first assumed a more likely scenario with an attacker's knowledge limited to observable demographics—sex and estimated age ("age_cohort" variable). The second modeled a highly informed adversary with access to all variables except the target ("hivstatus"). The synthetic data were generated from one round of real data, while control data came from a different round. For example, CTGAN models trained on R6 data were evaluated using R8 data as the control. We report results only where the attacker significantly outperformed both random guessing and the control baseline to ensure that measured risks reflected specific privacy leakage rather than general data utility in line with previous work using Anonymeter.[28]

Results

Fidelity

The univariate distributions between the real and synthetic data for the sex and pregnant variables for both R6 and R8 are shown in Figure 2 while the univariate distributions for all variables are presented in Supplementary File 1. The results show that CTGAN generated synthetic data that largely maintained the proportions of individual features between the real and synthetic data for both R6 and R8. For example, both the real and CTGAN generated datasets show a higher number of females than males, many respondents lacking a birth certificate, and widespread use of long-lasting

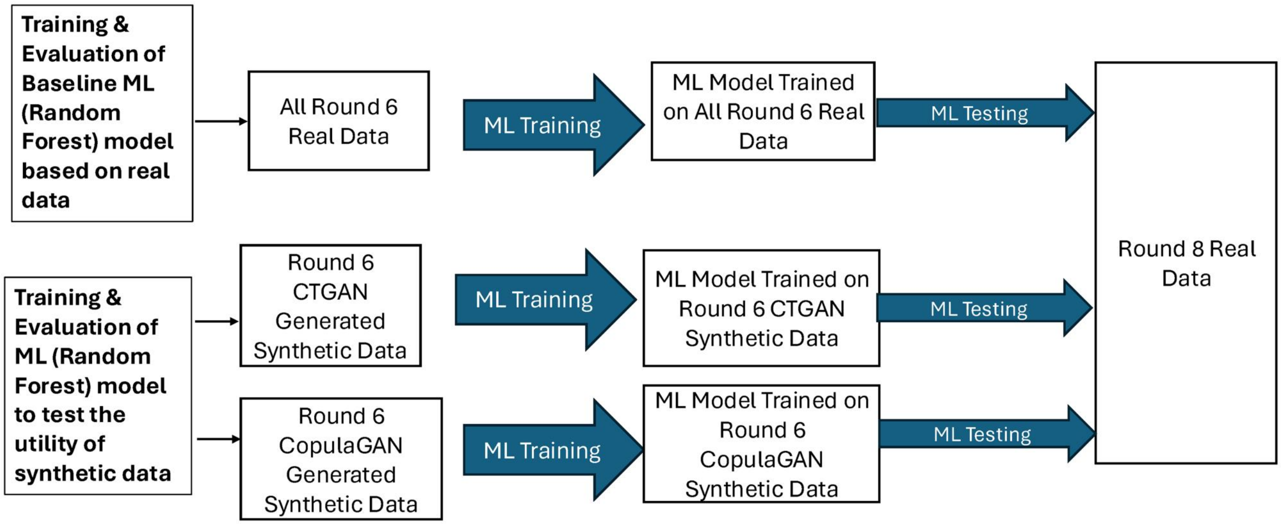
Training & Evaluation of Baseline ML (Random Forest) model based on real data
All Round 6 Real Data
ML Training
ML Model Trained on All Round 6 Real Data
ML Testing
Training & Evaluation of ML (Random Forest) model to test the utility of synthetic data
Round 6 CTGAN Generated Synthetic Data
ML Training
ML Model Trained on Round 6 CTGAN Synthetic Data
ML Testing
Round 6 CopulaGAN Generated Synthetic Data
ML Training
ML Model Trained on Round 6 CopulaGAN Synthetic Data
ML Testing
Round 8 Real Data

Figure 1. This figure illustrates how the experiements done for synthetic data utility test was setup.

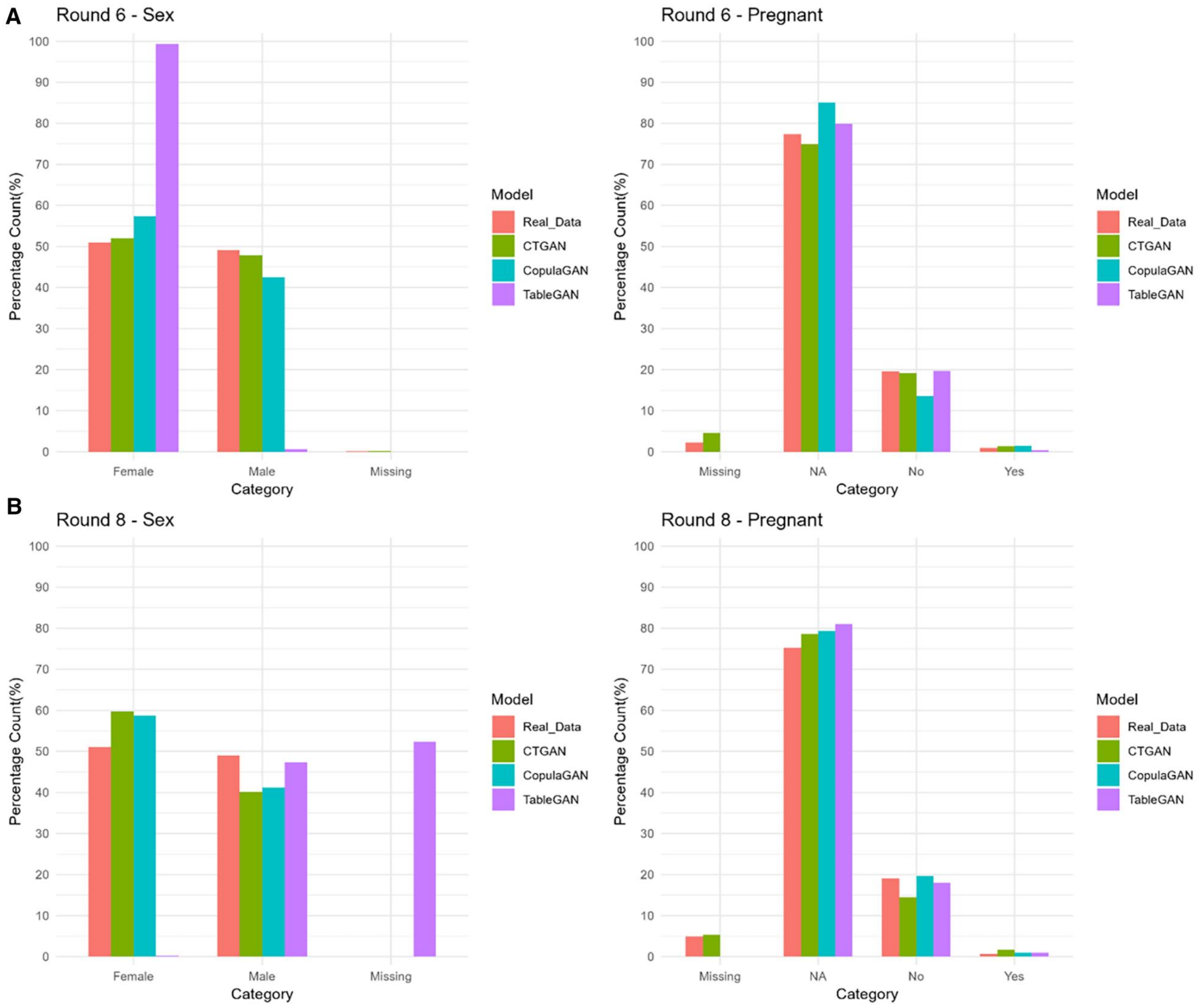

Figure 2. Univariate distributions. This figure provides the Percentage counts for the sex and pregnant variables in real and synthetically generated datasets where the graphs labelled in (A) represent round 6 variables (R6) and graphs in (B) represent round 8 (R8) variables. The percentage counts for all the variables is provided in the Supplementary Material.

insecticide-treated nets. Additionally, CTGAN effectively replicated missing values, reflecting those present in the original dataset. In contrast, TableGAN and CopulaGAN performed poorly in both rounds, failing to replicate the proportions of features and missing values. For instance, in the real data, the count for the sex variable in R6 was Female: 50.89%, whereas CTGAN generated Female: 51.99%, TableGAN generated Female: 99.37% and CopulaGAN produced Female: 57.34%. Similarly, for the pregnant variable, there were 2.27% missing values in the real data in R6, which was replicated by CTGAN, generating 4.62% missing values, while both TableGAN and CopulaGAN did not generate any blank values. Given the poor performance of TableGAN (eg, In R6, 99.37% of the data in TableGAN were female compared to 50.89% in the real data) coupled with its omission of an important variable (water treatment assessment-"watertre") we did not perform further analyses of TableGAN generated data as their utility would be extremely low.

To evaluate whether relationships between variables were preserved in synthetic data compared to the real dataset, we assessed bivariate distributions using joint frequency distribution. CTGAN-generated synthetic data that exhibited a higher degree of similarity to the real data compared to CopulaGAN. For instance, in both the real and CTGAN-generated R6 data, household heads were predominantly in the 25-59 age range (Real: 12.8%, CTGAN: 10.5%), while children were mainly in the 5-12 age range (Real: 18.5%, CTGAN: 18.3%) (Figure 3). In contrast, CopulaGAN failed to replicate these patterns as accurately. For example, for children the dominant age ranges were 12-59 months (10.2%) in CopulaGAN and 0-28 days (8.0%) (Figure 3).

We performed hierarchical clustering using the correlations between variables based on the computed associations between variable pairs to visualize the relationships between the real data and data generated by CTGAN and CopulaGAN. Variables clustered on the same branch in the real dataset also tended to cluster together on the CTGAN generated synthetic data (Figures 4 and 5). The heatmaps for both the synthetic data generated by CTGAN and real data were highly similar, though the synthetic dataset heatmap has a diminished intensity, reflecting weaker associations between variables in synthetic data versus real data (Figures 4 and 5).

When comparing the distribution of the pairwise associations (normalized mutual information) in the real and

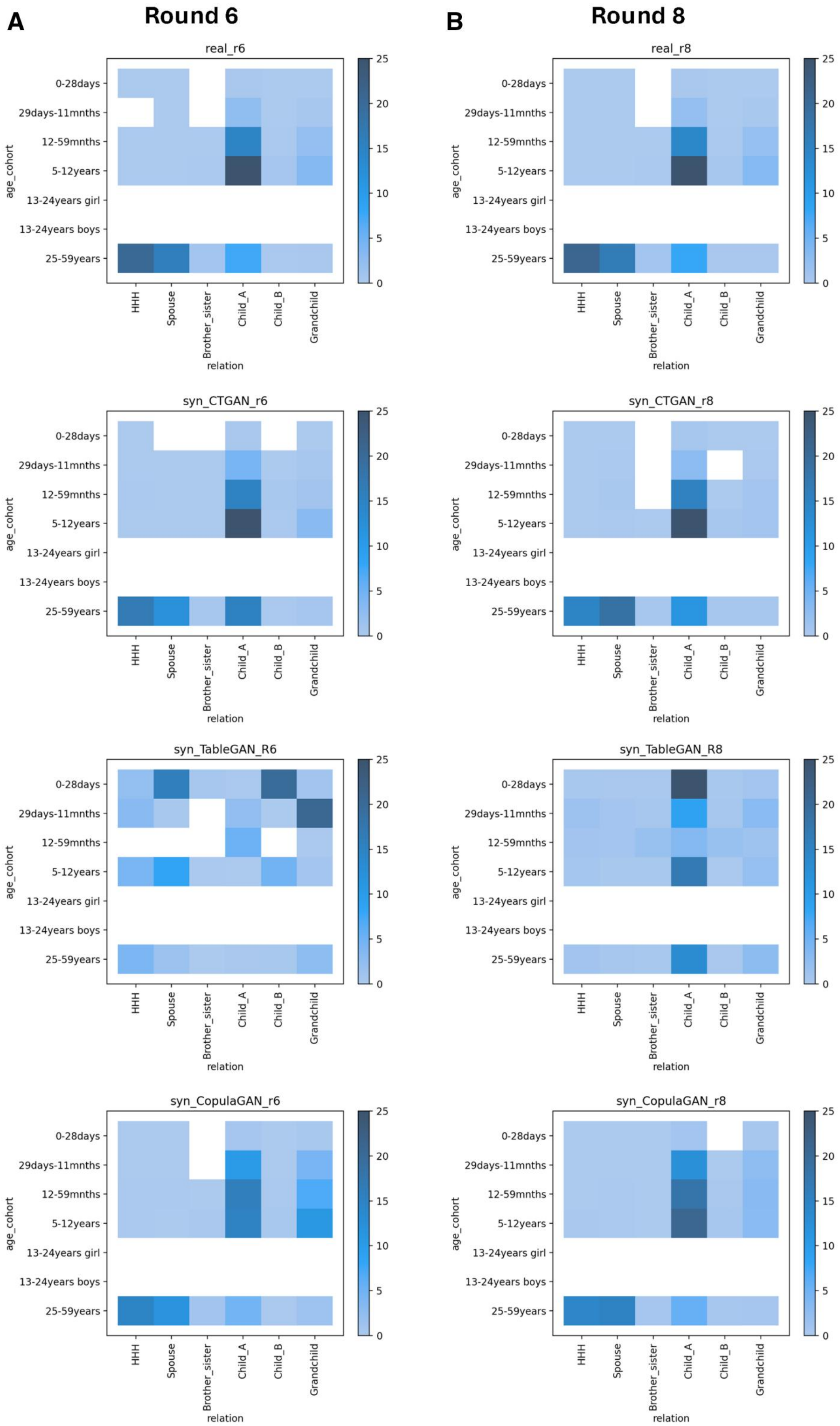

Figure 3. This figure provides the bivariate distributions for the relation and age_cohort variable. The distributions were computed for the real and synthetically generated data, comparing (A) round 6 and (B) round 8.

synthetic data, all the CTGAN and CopulaGAN models showed statistically significant differences from the real data in R6 and R8 with varying degrees of deviation from the real data based on the Kolmogorov-Smirnov tests (Table 1). CTGAN recorded the lowest D-values, as compared to CopulaGAN, indicating the least deviation from the real data (Table 1). The Wilcoxon rank-sum test revealed that synthetic data generated by CTGAN showed a significant difference in the magnitude of associations from the real data in R6, indicating some discrepancies (Table 1). However, the difference was not statistically significant in R8 suggesting the synthetic data closely resembled the real data but still

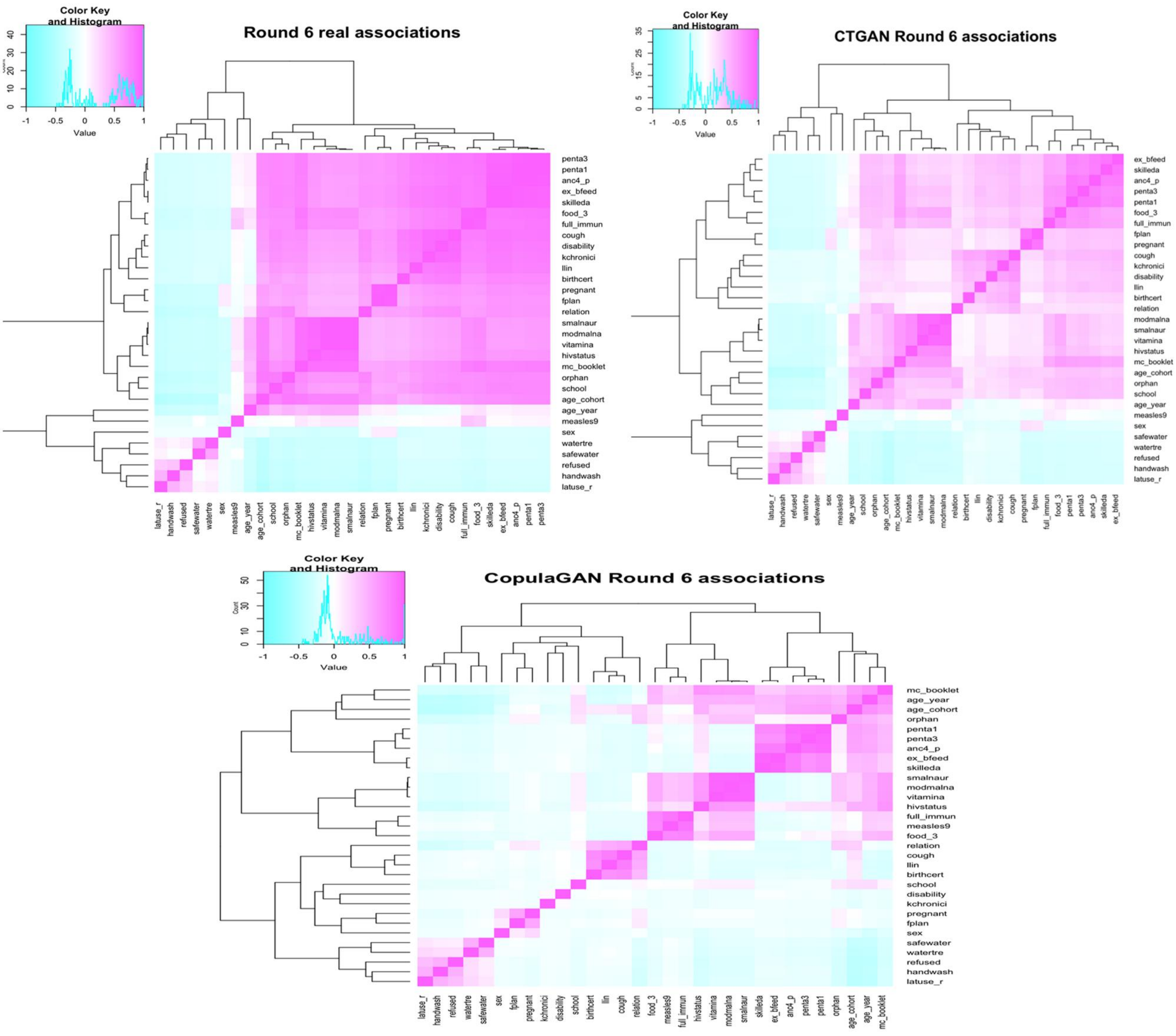

Figure 4. Heatmaps showing relationships between features in the real data and synthetic data generated by CTGAN and CopulaGAN for round 8. Heatmaps were constructed using the correlations between features based on their associations followed by hierarchical clustering or directly using the raw U associations between variable pair 6.

exhibited minor differences (Table S2). In contrast, synthetic data produced by CopulaGAN showed highly significant differences from the real data in both rounds, indicating larger discrepancies. Lastly, visual inspection of the CDF plots further supports these trends, with CTGAN generating synthetic data that is in closer alignment with the real data distribution (Figure S1).

Machine learning utility

The utility tests were consistent with the fidelity test findings, highlighting the strong performance of the CTGAN model across both rounds. When the RF model was trained on synthetic data from R8 and evaluated against real data from R6, its performance in predicting the target variable for functional latrine (latuse_r') varied. Both CTGAN and CopulaGAN showed results comparable to those achieved when the RF model was trained and tested on real data (Table 2). Notably, the RF model trained on synthetic data from R6 and evaluated against real data from R8 demonstrated enhanced performance, with CTGAN achieving high scores in accuracy and AUC-ROC, closely aligning with the performance of models using real data. While CopulaGAN also delivered strong results, CTGAN outperformed it in precision in some cases. These findings underscore the efficacy of CTGAN and CopulaGAN in producing synthetic data that preserves essential statistical properties for predictive analytics.

Privacy test

In the plausible attacker scenario, the risk was consistently low. For CTGAN, the risk remained 0.000 (95% CI, 0.000-0.000) in R6 and 0.019 (95% CI, 0.005-0.033) in R8. CopulaGAN showed a low risk of 0.066 (95% CI, 0.053-0.079) in R6 and 0.000 (95% CI, 0.000-0.000) in R8. In contrast, in the highly informed attack scenario, privacy risks varied by model and data round. For CTGAN, the risk was effectively zero in R6 but rose to 0.366 (95% CI, 0.343-0.388) in R8. CopulaGAN showed a similar pattern, with inconclusive results in R6, but a higher risk of 0.415 (95% CI, 0.397-0.433) in R8.

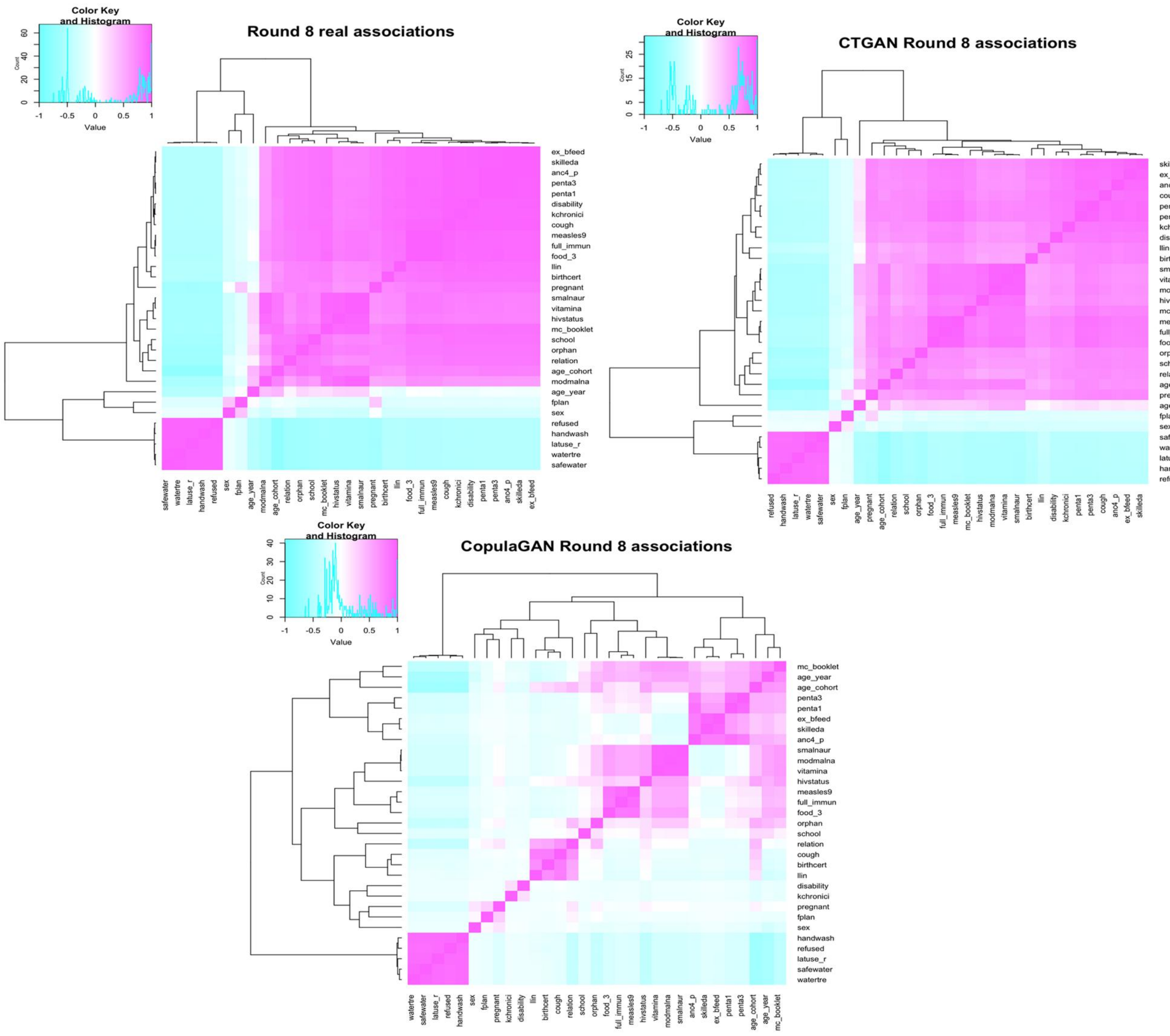

Figure 5. Heatmaps showing relationships between features in the real data and synthetic data generated by CTGAN and CopulaGAN for round 8. Heatmaps were constructed using the correlations between features based on their associations followed by hierarchical clustering or directly using the raw U associations between variable pairs.

Table 1. Statistical distribution of individual variables in synthetically generated data compared to the real data.

	Kolmogorov-Smirnov		Wilcoxon rank-sum	
Dataset	R6 D	R8 D	R6 Median	R8 Median
Real	NA	NA	0.17	0.27
CTGAN	0.22	0.23	0.12	0.23
CopulaGAN	0.32	0.37	0.03	0.03

D denotes the maximum absolute difference between the CDFs of variable pairs in the real data and synthetically generated data using the 3 GAN models.

Discussion

This study evaluates effectiveness of GAN models in generating high-quality tabular synthetic data based on HDSS data from a low- and middle-income country setting, as proof of concept. The CTGAN model generated the highest quality synthetic HDSS data, effectively preserving key trends and statistical properties of the original dataset while maintaining data utility and protecting individual privacy. This has significant implications for data sharing in LMICs, where research data is scarce, and existing data-sharing policies often impose strict limitations on access to data for research purposes. Synthetic data offer a viable solution by enabling broader data sharing without compromising participant confidentiality, fostering collaboration and innovation while ensuring ethical compliance and data security.

Synthetic data generation is an active area of research with many open questions and challenges. While there are no standard objective metrics to evaluate whether synthetic versions of a real dataset can replace real data, fidelity, utility, and privacy are currently the most common assessments.[29–31]

Consistent with the results obtained by Xu et al,[23] our analysis revealed that CTGAN significantly outperformed the other GAN models in generating synthetic data that closely mirrored the statistical properties of the real dataset, aligning with previous findings suggesting that CTGAN is particularly effective in capturing the complexities inherent in tabular data.[23] However, it is noteworthy that while CTGAN produced data that reflected the overall statistical characteristics

Table 2. Results of utility random forest (RF) classifier trained using real and synthetic R6 data and evaluated on real R8 data.

Dataset	Accuracy (%, 95% CI)	Precision (%, 95% CI)	Recall (%, 95% CI)	F1 score (%, 95% CI)	AUC-ROC (%, 95% CI)
Real Data	72.0 (71.7-72.3)	74.2 (73.9-74.4)	72.0 (71.7-72.3)	68.3 (68.1-68.6)	88.8 (88.6-89.0)
CTGAN Data	72.4 (72.1-72.7)	71.8 (71.5-72.1)	72.4 (72.1-72.7)	71.4 (71.1-71.7)	77.4 (77.1-77.7)
CopulaGAN Data	70.7 (70.4-71.0)	72.2 (71.9-72.5)	70.7 (70.4-71.0)	66.8 (66.5-67.1)	76.3 (76.0-76.6)

of the real dataset, the pairwise associations within the CTGAN-generated data were generally weaker compared to those found in the real data. This suggests that, although CTGAN excels in replicating individual variable distributions, it may not fully capture the strength of relationships between variables. In contrast, both CopulaGAN and TableGAN demonstrated poor performance in generating synthetic data. This limitation highlights the challenges associated with using certain models for specific data types and emphasizes the importance of model selection in the synthetic data generation process. Overall, while CTGAN shows promise, further work is needed to enhance the fidelity of pairwise associations to achieve a more comprehensive representation of the underlying relationships in the real dataset.

The experiments conducted underscored the importance of clearly defining the purpose of synthetic data generation before embarking on generation, as this guides the choice of approach and validation techniques. For instance, if the synthetic data are intended for a classification task, validation should involve comparing the performance of models trained on synthetic data and evaluated on real data against those trained and evaluated on original data. In this study, we assessed the utility of synthetic data by training a Random Forest model on synthetic data and evaluating it on real data, then comparing the results to when the model is trained and tested on real data. The utility test results aligned with the fidelity outcomes, highlighting the robustness of the CTGAN model across both rounds. The Random Forest model trained on synthetic data from R8 and evaluated on real data from R6 showed varied performance in predicting the target variable “latuse_r.” Both CTGAN and CopulaGAN produced results comparable to real data models, with CopulaGAN outperforming CTGAN in most metrics except precision, where CTGAN excelled. Conversely, the Random Forest model trained on CTGAN synthetic data from R8 demonstrated the highest overall performance, except in precision, where CopulaGAN was superior. These findings illustrate the specific strengths of each synthetic data generation model in supporting predictive analytics.

Synthetic data generation does not necessarily provide perfect privacy guarantees. Models used to generate synthetic data can be “tricked” into reconstructing some of the real data instances.[32] Although the KRHDSS training dataset contains limited sensitive personal information, the privacy tests done in this study illustrate how privacy risks in synthetic data can still emerge depending on the attack scenario and data properties. Under a highly informed attack scenario, low to moderately high privacy risks were observed, with R8 consistently showing greater exposure than R6—suggesting that longitudinal variation and additional information over time may heighten vulnerability. By contrast, the plausible attack scenario revealed consistently low privacy risk across both models and rounds. These results underscore the importance of aligning privacy assessments with realistic threat models and recognizing how evolving dataset characteristics, such as those across release rounds, can influence risk. Responsible deployment of synthetic datasets should therefore include targeted privacy assessments that account for both attacker knowledge and the potential accumulation of risk across longitudinal releases.

The limitations of this study are as follows: (1) Generalizability concerns, as the applicability of this work beyond the specific Kenyan dataset (KRHDSS) used is uncertain. Additionally, the limited training data from only 2 rural sub-counties may restrict the generalization of findings. Expanding the study to include urban or international datasets, such as those examined in Ghosheh et al[33] could provide more comprehensive insights. (2) Additional assessment of privacy will be needed as ascertainment of perfect privacy guarantees depend on the nature of attacks and sophistication of the attacker which are beyond the scope of this work. (3) This study focused exclusively on evaluating GAN-based models for synthetic data generation. Conducting additional experiments with other models using an HDSS dataset from an LMIC resource setting could offer deeper insights into the effectiveness of different approaches for synthetic data generation.

Future research studies can seek to explore the utility and effectiveness of CTGAN for synthetic data generation given other tabular data that contain mixed data types and compare to other models that have so far been reported in literature such as the tabular transformer generative adversarial network (TT-GAN).[34] Additionally, future research in synthetic data generation should address the ethical, legal, and social implications (ELSI) associated with its use.[35] Currently, synthetic data generation is not directly addressed by data privacy laws, largely because policy changes fail to keep pace with rapid technological progress.[36] This leads to confusion, leaving room for different legal interpretations and litigation. Similar to de-identified data—which remains more directly linked to participants—synthetic data generated from research participants may be used in ways against their wishes. For example, broadly shared patient synthetic data may be used commercially without permission of the patients or benefit sharing; data may also be used to generate inferences about groups that individual patients may find problematic. Due to these collective concerns, a careful assessment of ELSI of synthetic data is needed on a case-by-case basis before broad sharing. Additionally, future studies might explore the generation of synthetic longitudinal data from an HDSS dataset, considering that HDSS data involve repeated measures of the same participants over time.

Conclusion

We demonstrated the use of the CTGAN technique to generate high-quality synthetic data from a HDSS in an LMIC setting as proof-of-principle. The CTGAN approach was superior in generating synthetic data that maintained the statistical properties available in the real data while maintaining utility, data privacy, and confidentiality. Synthetic data generation offers a potential avenue to address privacy, regulatory, and legal concerns around data sharing and enhance the broad utilization of valuable datasets, particularly in LMICs where a growing repository of HDSS data form a critical resource for research, policy-making, and public health interventions.

Acknowledgments

We would like to thank The Aga Khan University for authorizing the use and access of the KRHDSS dataset that was used to train the CTGAN model for synthetic data generation.

Author contributions

Dorcas G. Mwigereri (Conceptualization, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing), Nigel T. Kamotho (Conceptualization, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing), Akbar K. Waljee (Funding acquisition, Methodology, Supervision, Writing—review & editing), Eileen M. Weinheimer-Haus (Project administration, Supervision, Writing—review & editing), Ryan T. Rego (Conceptualization, Data curation, Methodology), Farhana Alarakhiya (Conceptualization, Funding acquisition, Supervision), Anthony K. Ngugi (Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing—original draft, Writing—review & editing), Nicholson Price II (Methodology, Writing—review & editing), Ji Zhu (Investigation, Methodology, Supervision, Writing—review & editing), Stephen Peter Wong (Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing), and Geoffrey H. Siwo (Conceptualization, Formal analysis, Methodology, Supervision, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the Office of The Director, National Institutes of Health (OD), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the National Institute of Mental Health (NIMH), the Fogarty International Center (FIC) of the National Institutes of Health under award number U54TW012089 (Abubakar A and Waljee AK) and 3U54TW012089-03S1, and the Novo Nordisk Foundation under award number NNF23SA0087056 (Price WN). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest

All authors have no competing interest to declare.

Data availability

Project materials, including code and documentation, are available at https://osf.io/2ptdq/overview. The synthetic dataset generated in this study using CTGAN is publicly available and can be accessed at https://www.icpsr.umich.edu/web/ICPSR/studies/39209.

References

1. Adazu K, White M, Findley S, Collinson M. *The Dynamics of Migration, Health and Livelihoods: INDEPTH Network Perspectives*. Ashgate Publishing; 2010.
2. Byass P, Berhane Y, Emmelin A, et al. The role of Demographic Surveillance Systems (DSS) in assessing the health of communities: an example from rural Ethiopia. *Public Health*. 2002;116:145-150. https://doi.org/10.1038/sj.ph.1900837
3. Kaewkungwal J, Adams P, Sattabongkot J, Lie RK, Wendler D. Issues and challenges associated with data-sharing in LMICs: perspectives of researchers in Thailand. *Am J Trop Med Hyg*. 2020;103:528-536. https://doi.org/10.4269/ajtmh.19-0651
4. Hinga AN, Molyneux S, Marsh V. Towards an appropriate ethics framework for health and demographic surveillance systems (HDSS): learning from issues faced in diverse HDSS in Sub-Saharan Africa. *BMJ Glob Health*. 2021;6:e004008. https://doi.org/10.1136/bmjgh-2020-004008
5. Templ M, Kanjala C, Siems I. Privacy of study participants in open-access health and demographic surveillance system data: requirements analysis for data anonymization. *JMIR Public Health Surveill*. 2022;8:e34472. https://doi.org/10.2196/34472
6. Bak M, Madai VI, Fritzsche MC, Mayrhofer MT, McLennan S. You can't have AI both ways: balancing health data privacy and access fairly. *Front Genet*. 2022;13:929453. https://doi.org/10.3389/fgene.2022.929453
7. Munung NS, Staunton C, Mazibuko O, Wall PJ, Wonkam A. Data protection legislation in Africa and pathways for enhancing compliance in big data health research. *Health Res Policy Syst*. 2024;22:145. https://doi.org/10.1186/s12961-024-01230-7
8. Hähnel M. Ethical challenges and solutions in AI-driven medical data management: a focus on distributed machine learning. *Discov Artif Intell*. 2025;5:53. https://doi.org/10.1007/s44163-025-00266-0
9. Kuo NI-H, Garcia F, Sönnerborg A, et al.; EuResist Network Study Group. Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: example using antiretroviral therapy for HIV. *J Biomed Inform*. 2023;144:104436. https://doi.org/https://doi.org/10.1016/j.jbi.2023.104436
10. Pacheco F, Hermosilla G, Piña O, et al. Generation of synthetic data for the analysis of the physical stability of tailing dams through artificial intelligence. *Mathematics*. 2022;10:4396. https://doi.org/10.3390/math10234396
11. Klein P, Bergmann R. Data generation with a physical model to support machine learning research for predictive maintenance. *CEUR Workshop Proc*. 2018;2191:179-190.
12. Nečas D, Klapetek P. Synthetic data in quantitative scanning probe microscopy. *Nanomaterials (Basel)*. 2021;11:1746. https://doi.org/10.3390/nano11071746
13. Devaux SE, Wehmeyer C. An overview of synthetic data types and generation methods. *KDnuggets News*. Accessed May 7, 2023. https://www.kdnuggets.com/2021/02/overview-synthetic-data-types-generation-methods.html
14. Ho, J Chen, X Srinivas, A Duan, Y, Abbeel, P. Flow++: improving flow-based generative models with variational dequantization and architecture design. In: *36th International Conference on Machine Learning, ICML 2019*. 2019: 4827–4842.
15. Razghandi, M Zhou, H Erol-Kantarci, M, Turgut, D. Variational autoencoder generative adversarial network for synthetic data generation in smart home. In: *ICC 2022—IEEE International*

Conference on Communications. 2022: 4781–4786. https://doi.org/10.1109/ICC45855.2022.9839249

16. Sivakumar J, Ramamurthy K, Radhakrishnan M, Won D. Synthetic sampling from small datasets: a modified mega-trend diffusion approach using k-nearest neighbors. *Knowl Based Syst*. 2022;236:107687. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107687.
17. Coutinho-Almeida, J Rodrigues, PP, Cruz-Correia, RJ. GANs for tabular healthcare data generation: a review on utility and privacy. In: Soares C, Torgo L, eds. *Discovery Science*. Cham: Springer International Publishing; 2021:282-291.
18. Rujas M, Martín Gómez del Moral Herranz R, Fico G, Merino-Barbancho B. Synthetic data generation in healthcare: a scoping review of reviews on domains, motivations, and future applications. *Int J Med Inform*. 2025;195:105763. https://doi.org/10.1016/j.ijmedinf.2024.105763
19. Chen H, Jajodia S, Liu J, Park N, Sokolov V, Subrahmanian VS. Faketables: using GANs to generate functional dependency preserving tables with bounded real data. In: *IJCAI International Joint Conference on Artificial Intelligence*. 2019: 2074–2080. https://doi.org/10.24963/ijcai.2019/287
20. Zhao Z, Kunar A, Van der Scheer H, Birke R, Chen LY. *CTAB-GAN: Effective Table Data Synthesizing*. Vol 1. Association for Computing Machinery; 2021.
21. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Doshi-Velez F, Fackler J, Kale D, Ranganath R, Wallace B, Wiens J, eds. *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Vol 68. PMLR; 2017: 286–305.
22. CopulaGANSynthesizer. Accessed August 21, 2024. https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/copulagansynthesizer
23. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: *Advances in Neural Information Processing Systems*, Vol. *32*. NeurIPS; 2019.
24. Ngugi AK, Walraven G, Orwa J, Lusambili A, Kimani M, Luchters S. Community-driven data revolution is feasible in developing countries: experiences from an integrated health information and surveillance system in Kenya. *J Glob Health Rep*. 2021;5. https://doi.org/10.29392/001c.25977
25. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. *Proceedings VLDB Endowment*. 2018;11:1071-1083. https://doi.org/10.14778/3231751.3231757
26. Welcome to the SDV!! Synthetic data vault. Accessed May 25, 2025. https://docs.sdv.dev/sdv
27. CopulaGAN model. Accessed August 27, 2024. https://sdv.dev/SDV/user_guides/single_table/copulagan.html
28. Giomi M, Boenisch F, Wehmeyer C, Tasnádi B. A unified framework for quantifying privacy risk in synthetic data. *PoPETs*. 2023;2023:312-328. https://doi.org/10.56553/popets-2023-0055
29. Xia Y, Wang C.-H, Mabry J, Cheng G. 2024. Advancing Retail Data Science: Comprehensive Evaluation of Synthetic Data. https://arxiv.org/abs/2406.13130
30. Yuan Y, Liu Y, Cheng L. 2025. A Multi-Faceted Evaluation Framework for Assessing Synthetic Data Generated by Large Language Models. https://arxiv.org/abs/2404.14445
31. Takyar A. Synthetic data: types, generation, evaluation, use cases and applications. Accessed August 27, 2024. https://www.leewayhertz.com/what-is-synthetic-data/#How-to-evaluate-synthetic-data-quality
32. Jordon J, et al. Hide-and-seek privacy challenge: synthetic data generation vs patient re-identification. In: Escalante HJ and Hofmann K, eds. *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*. Vol 133. PMLR. 2021:206–215.
33. Ghosheh GO, Thwaites CL, Zhu T. Synthesizing electronic health records for predictive models in low-Middle-income countries (lmics). *Biomedicines*. 2023;11:1749.https://doi.org/10.3390/biomedicines11061749
34. Kang HYJ, Ko M, Ryu KS. Tabular transformer generative adversarial network for heterogeneous distribution in healthcare. *Sci Rep*. 2025;15:10254. https://doi.org/10.1038/s41598-025-93077-3
35. Susser D, Schiff DS, Gerke S, et al. Synthetic health data: real ethical promise and peril. *Hastings Cent Rep*. 2024;54:8-13. https://doi.org/https://doi.org/10.1002/hast.4911.
36. Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. *Lancet*. 2022;399:1601-1602. https://doi.org/10.1016/S0140-6736(22)00232-X

JAMIA Open, 2025, 8, 1–10
https://doi.org/10.1093/jamiaopen/ooaf137
Research and Applications