Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Synthetic tabular healthcare data in health informatics: fixed-size substitution and incremental-value controls for predictive utility Application Paper

Miguel Álvaro Andrade Oliveira

miguelalvaro13@gmail.com

Faculdade de Medicina da Universidade do Porto

Miguel Ricou

Faculdade de Medicina da Universidade do Porto

Alberto Freitas

Faculdade de Medicina da Universidade do Porto

Research Article

Keywords: healthcare data, machine learning evaluation, predictive utility, synthetic data, tabular data

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11128489/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Adding synthetic records can make a predictive model appear better simply because it receives more training rows—not because synthetic data add useful information. Repeated predictor profiles split across training and test sets can create a second source of overoptimism. We designed a benchmark that controls both problems. Using three public diabetes-related healthcare datasets, we evaluated Gaussian Copula, Conditional Tabular Generative Adversarial Network (CTGAN), and Tabular Variational Autoencoder (TVAE) with logistic regression, random forest, and histogram gradient boosting across 10 paired replications. First, we held classifier-training size constant while replacing 0% to 100% of real rows with synthetic rows. Second, we compared N real rows (RR-N), N real plus N bootstrap-real rows (RB-2N), and N real plus N synthetic rows (RSADD-2N). Exact duplicate predictor profiles were kept within a single data partition. At fixed size, synthetic substitution produced significantly lower area under the receiver operating characteristic curve (AUC) in 89 of 108 model-specific contrasts after Holm correction, significantly higher AUC in four, and no significant difference in 15. In the additive comparison, synthetic augmentation was significantly worse in 21 of 27 contrasts against RR-N and 16 of 27 against RB-2N. Favourable effects occurred, but were isolated rather than systematic. Under the evaluated conditions, synthetic data therefore offered no general predictive-utility benefit. The study provides a practical framework for separating synthetic-data value from sample-size effects and partition leakage. A reproducibility package supplies code, configurations, public inputs, and validated outputs. These findings do not establish privacy, fairness, clinical equivalence, or deployment suitability.

1. Introduction

Health informatics teams need realistic data for predictive modelling, software testing and workflow comparisons, but patient records are sensitive and hard to share. Synthetic generators learn from real records to create rows intended to preserve useful structure without releasing source records [1, 2].

Resemblance alone is insufficient. Healthcare evaluations assess predictive utility, distributional similarity and disclosure risk, but heterogeneous designs do not establish a graded fixed-size substitution benchmark [2–6]. Recent application studies have reported higher classification accuracy after combining real and synthetic data, but when the augmented arm contains more rows than the real-only comparator, the result cannot isolate synthetic-data contribution from training-size effects [7].

We evaluate a leakage-controlled framework that combines a graded fixed-size substitution curve with a protocol-locked incremental-value control on unseen real test data. Its applied contribution is to separate the effect of synthetic-data composition from the effect of enlarging the classifier-training set, rather than to propose a new generator.

2. Objectives

We tested whether replacing real training observations with synthetic observations preserved or improved prespecified area under the receiver operating characteristic curve (AUC) at fixed N, and

whether adding N synthetic rows to all N real rows outperformed RR-N and an equal-size bootstrap-real 2N control. Secondary outcomes covered utility, fidelity, privacy-risk warning signs and subgroup stability.

3. Methods

3.1 Study design and datasets

We used three public binary-outcome datasets: CDC Diabetes Health Indicators (CDC) (253,680 rows; prevalence 13.9%), Diabetes 130-US Hospitals (101,766 rows; prevalence 11.2%) and Early Stage Diabetes Risk Prediction (520 rows; prevalence 61.5%).[8–11] In Early Stage, 520 rows represented 251 exact predictor signatures; there were 269 duplicate-excess rows, 107 repeated groups and maximum multiplicity 16. No signature had conflicting outcome labels. The datasets concern diabetes but represent population-health risk factors, hospital records and a clinical questionnaire.

Each dataset used a fixed approximately 60/20/20 train/validation/test split, stratified where feasible. Diabetes 130-US patient_nbr groups stayed together. In Early Stage, all 16 predictors, excluding the outcome, defined exact-profile groups; stratified group splitting kept each signature in one partition, with a post-split assertion of zero overlap. Generators learned only from real training data; outcomes used held-out real test data. Seed 42 defined the partition reused across replications, so uncertainty is conditional on one split.

3.2 Generators, models and scenarios

We tested Gaussian Copula, Conditional Tabular Generative Adversarial Network (CTGAN) and Tabular Variational Autoencoder (TVAE).12,13 Synthetic pools matched real training size. Logistic Regression, Random Forest and HistGradientBoostingClassifier used prespecified settings. Ten replication-specific seeds varied generation, scenario sampling and stochastic fitting; partitions, hyperparameters and pool size stayed fixed. Online Resource 1 documents software, parameters and seeds.

N denotes real training size. In substitution, RR used N real rows; RS25, RS50 and RS75 replaced 25%, 50% and 75% with synthetic rows; SR used N synthetic rows. Each generator learned from all N real training rows and produced one N-row pool per replication. The control compared RR-N, RB-2N (N real plus N target-stratified bootstrap-real rows) and RSADD-2N (N real plus N generator-derived rows). Compositions were reused across classifiers. Training size was fixed in substitution and matched at 2N between additive arms; generator exposure was not equalised. All arms used unchanged real test records. Supplementary Fig. S4.1 summarises the three-arm design and seven prespecified AUC contrasts.

3.3 Evaluation and aggregation

AUC was primary. Area under the precision–recall curve (AUPRC), F1, balanced accuracy and Brier score were secondary inferential outcomes; recall and precision were descriptive, outside paired-test families.

[14–16] Results were paired by replication. Model-averaged percentile 95% intervals resampled the 10 within-replication model means with replacement 2,000 times (seed 42). Two-sided paired Wilcoxon tests used 10 pairs. Holm correction applied within dataset × model × metric across 12 fixed-size contrasts; the incremental control used seven. Paired rank-biserial correlation and Cliff's delta supplemented p-values. The 0.50 threshold was fixed; validation supported diagnostics, not model selection or threshold optimization.

Fidelity covered marginal distributions, numeric distances, feature/correlation gaps and target prevalence. Privacy-risk proxies measured exact and near matches and proximity to training versus holdout records; these are warnings, not anonymity or membership probabilities. Subgroup stability used available sex/gender, age and race/ethnicity fields, excluding groups with fewer than 30 test records or one outcome class. Online Resource 2 provides definitions, denominators and tables.

3.4 Duplicate-safe partitioning and analysis provenance

A pre-submission audit found Early Stage leakage from row-wise splitting: 85 predictor signatures crossed partitions (53 train–validation, 51 train–test and 29 validation–test). We invalidated affected estimates, grouped splits by predictor signature and reran all affected primary and incremental slices. The corrected split retained 312/104/104 rows, prevalence 0.6154 in each partition and zero overlap. CDC and Diabetes 130-US analyses were retained unchanged. This post-lock amendment corrected data integrity without changing planned arms, outcomes or contrasts.

To support transparent evaluation and independent checking, Online Resource 5 supplies executable code, locked configurations, automated tests, public input data, validation scripts, and validated outputs for both the fixed-size benchmark and the incremental-value control. SHA-256 checksums and a file manifest support integrity checking. Full rerun commands are included; complete reruns may require substantial computing time.

3.5 Ethical considerations

The study used public benchmarks without participant recruitment, contact, intervention, animal research or author collection of identifiable private information.

3.6 Use of AI-assisted tools

AI-assisted tools were used under human supervision for, synthetic-data generation code and analytical checks. The authors retain full responsibility for the work.

4. Results

At fixed N, the mean AUC difference relative to real-only training (ΔAUC) across dataset–generator conditions was − 0.0016 at RS25, − 0.0108 at RS50, − 0.0323 at RS75 and − 0.1469 at SR. Of 108 model-specific AUC contrasts, 89 were significantly worse after Holm, four significantly better and 15 non-significant (Table 1; Fig. 1). Corrected Early Stage RR mean AUC was 0.9375 for HistGradientBoosting,

0.9320 for Logistic Regression and 0.9782 for Random Forest; corresponding legacy estimates were 0.9969, 0.9875 and 0.9933, confirming inflation from leakage.

Table 1
Descriptive, post-selection fixed-size predictive-utility results

Dataset	Synthetic share	Best generator	Mean ΔAUC (95% CI)	Interpretation
CDC Diabetes Health Indicators	25%	CTGAN	−0.0018 (− 0.0021 to − 0.0015)	Loss vs RR
CDC Diabetes Health Indicators	100%	CTGAN	−0.0214 (− 0.0245 to − 0.0188)	Loss vs RR
Diabetes 130-US Hospitals	25%	Gaussian Copula	−0.0060 (− 0.0070 to − 0.0048)	Loss vs RR
Diabetes 130-US Hospitals	100%	Gaussian Copula	−0.0578 (− 0.0589 to − 0.0566)	Loss vs RR
Early Stage Diabetes	25%	Gaussian Copula	+ 0.0128 (+ 0.0086 to + 0.0166)	Gain vs RR
Early Stage Diabetes	100%	TVAE	−0.1209 (− 0.1460 to − 0.0940)	Loss vs RR

Note: Post-selection best-generator results. Intervals resampled 10 replication-level means 2,000 times with replacement (seed 42), without selection adjustment. All training sets had N rows; see Online Resource 2.

In the incremental control, RB-2N versus RR-N model-averaged ΔAUC was − 0.0028 for CDC, − 0.0057 for Diabetes 130-US and + 0.0046 for Early Stage; six of nine model-specific contrasts were significantly worse, one better and two non-significant. Eight of nine dataset–generator RSADD-2N versus RR-N means were negative; Gaussian Copula on Early Stage was positive (+ 0.0155). At model level, 21/27 contrasts were significantly worse, two better and four non-significant. Versus RB-2N, 16/27 were worse, two better and nine non-significant; positive model-averaged effects occurred for Gaussian Copula on Diabetes 130-US (+ 0.0010) and Early Stage (+ 0.0109). Thus favourable effects were context-specific rather than systematic. Supplementary Fig. S4.1 displays the corresponding model-averaged paired ΔAUC estimates and 95% intervals.

In the fixed-size benchmark, 13 F1 and 13 balanced-accuracy contrasts improved significantly; 53 and 57 worsened, respectively. In the additive control, RSADD versus RR yielded four significantly better and 12 worse contrasts for each metric; versus RB, three improved while 16 F1 and 15 balanced-accuracy contrasts worsened. These fixed-threshold shifts are not evidence of better discrimination.

Across fixed-size conditions, TVAE and Gaussian Copula had similar overall mean ΔAUC losses (− 0.0329 and − 0.0332), versus − 0.0777 for CTGAN. Gaussian Copula had the smallest mean Hellinger distance (0.056 versus 0.100 for CTGAN and 0.164 for TVAE). Mean near-duplicate rates were 0.0099,

0.0211 and 0.2341, respectively; relative-proximity proxy rates were 0.731, 0.749 and 0.774. These structure- and partition-sensitive proxies do not establish anonymity.

Across 13,320 eligible subgroup–replication contrasts, median ΔAUC was − 0.0158 (IQR − 0.0450 to − 0.0034), 83.7% were negative and subgroup denominators ranged from 32 to 28,531. Scenario minima were − 0.278, − 0.347, − 0.486 and − 0.950 from RS25 to SR; these are single-replication warnings, not average effects or fairness certification. Rankings remain conditional on the evaluated datasets, models, split and mixture design.

5. Discussion

5.1 Principal findings

At fixed N, most significant AUC effects were adverse, but four favourable contrasts preclude an absolute no-benefit claim. The added control showed no systematic advantage for bootstrap-real doubling or synthetic addition. Isolated gains indicate context-specific value, not a general augmentation rule or single causal explanation for the historical 2N result. These findings show why predictive gains reported after enlarging training sets with synthetic rows should be interpreted against equal-size real-data controls [7]. Duplicate-safe reanalysis reduced Early Stage held-out discrimination and changed aggregate counts, showing why repeated predictor profiles must remain within partitions.

5.2 Generator and metric interpretation

No generator was universally superior. CTGAN led CDC AUC; Gaussian Copula led Diabetes 130-US and corrected Early Stage at RS25; TVAE led corrected Early Stage at RS50, RS75 and SR. Threshold-dependent gains without AUC gain require calibration and decision-context analysis rather than being labelled better discrimination.

5.3 Clinical implications and supplementary evidence

Low-loss synthetic mixtures may support teaching and software testing when task-specific degradation is tolerable; clinical use requires real-data validation, calibration, subgroup review and governance. Online Resources 2 and 4 summarise results from 30 primary slices (1,170 metrics, 1,080 matched deltas, 540 tests) and 30 incremental slices (450 metrics, 315 complete paired tests), respectively, with no recorded failures.

6. Limitations

Evidence covers three diabetes-related public datasets, three generators and one fixed split. Ten replications measure stochastic variability within that split, not population or alternative-split variability. The minimum two-sided Wilcoxon p-value is 0.001953; minimum Holm-adjusted values are 0.0234 for 12 fixed-size contrasts and 0.013672 for seven incremental contrasts. RB-2N controls repeated

presentations and implicit weighting, not N new independent patients, and cannot identify every mechanism behind the historical 2N result.

Each generator learned from all N real training records, so fixed classifier-training size does not equalise informational exposure. Historical TVAE fallback and target-support repair counts could not be independently verified from the supplied records. Grouping prevents exact predictor-profile leakage in Early Stage but cannot make its small, highly duplicated sample representative. Privacy measures were screening proxies, not adversarial attacks; relative-proximity rates are not membership probabilities. Subgroup definitions varied; small or single-class groups were excluded. Neither fairness nor clinical validity was established.

7. Conclusions

Synthetic substitution and addition produced no systematic AUC benefit, although isolated model–dataset pairs improved. Incremental value is context-dependent, not absent. These controls do not establish a single explanation for the historical 2N result. Duplicate-safe splitting, real-data validation, calibration, subgroup and privacy checks remain necessary.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

The study used public benchmarks without participant recruitment, contact, intervention, animal research or author collection of identifiable private information.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

No specific funding was received.

Author Contribution

M.A.A.O. contributed to conceptualization, methodology, and writing the original draft. M.R. contributed to conceptualization, formal analysis, and review and editing of the manuscript. A.F. contributed to supervision, methodology, and review and editing of the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors thank colleagues for helpful discussions and acknowledge the institutional research environment.

Data Availability

The datasets used in this study are publicly available from the sources cited in the manuscript. The public input files used in the analyses, together with supporting materials and validated outputs, are provided in Online Resource 5 (reproducibility package). Additional tabulated results are provided in Online Resources 2 and 4.

Code availability

Executable source code, locked configurations, automated tests, validation scripts, public input data and validated outputs for the fixed-size and incremental analyses are provided in Online Resource 5. SHA-256 checksums and a file manifest support integrity checking. No external API credentials are required; the LLM component is a mock scaffold excluded from empirical claims.

References

1. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng. 5, 493–497 (2021). https://doi.org/10.1038/s41551-021-00751-8

2. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., Sales, A.P.: Generation and evaluation of synthetic patient data. BMC Med. Res. Methodol. 20, 108 (2020). https://doi.org/10.1186/s12874-020-00977-1

3. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.P.: Generation and evaluation of privacy preserving synthetic health data. Neurocomputing 416, 244–255 (2020). https://doi.org/10.1016/j.neucom.2019.12.136

4. Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., et al.: Privacy-preserving generative deep neural networks support clinical data sharing. Circ. Cardiovasc. Qual. Outcomes 12, e005122 (2019). https://doi.org/10.1161/CIRCOUTCOMES.118.005122

5. Tucker, A., Wang, Z., Rotalinti, Y., Myles, P.: Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit. Med. 3, 147 (2020). https://doi.org/10.1038/s41746-020-00353-9

6. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using generative adversarial networks. Proc. Mach. Learn. Res. 68, 286–305 (2017)

7. Ahmed, H.A., Nepomuceno, J.A., Vega-Márquez, B., Nepomuceno-Chamorro, I.A.: Synthetic data generation for healthcare: exploring generative adversarial networks variants for medical tabular data. Int. J. Data Sci. Anal. 20, 5739–5754 (2025). https://doi.org/10.1007/s41060-025-00816-w

8. UCI Machine Learning Repository: CDC Diabetes Health Indicators [data set] (2017). https://doi.org/10.24432/C53919

9. Strack, B., DeShazo, J.P., Gennings, C., et al.: Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed. Res. Int. 2014, 781670 (2014). https://doi.org/10.1155/2014/781670

10. Clore, J., Cios, K., DeShazo, J., Strack, B.: Diabetes 130-US Hospitals for Years 1999–2008 [data set]. UCI Machine Learning Repository (2014). https://doi.org/10.24432/C5230J

11. UCI Machine Learning Repository: Early Stage Diabetes Risk Prediction [data set] (2020). https://doi.org/10.24432/C5VG8H

12. Patki, N., Wedge, R., Veeramachaneni, K.: The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. IEEE (2016). https://doi.org/10.1109/DSAA.2016.49

13. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. Adv. Neural Inf. Process. Syst. 32 (2019). https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

14. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010

15. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10, e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432

16. Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3 (1950)

Figures

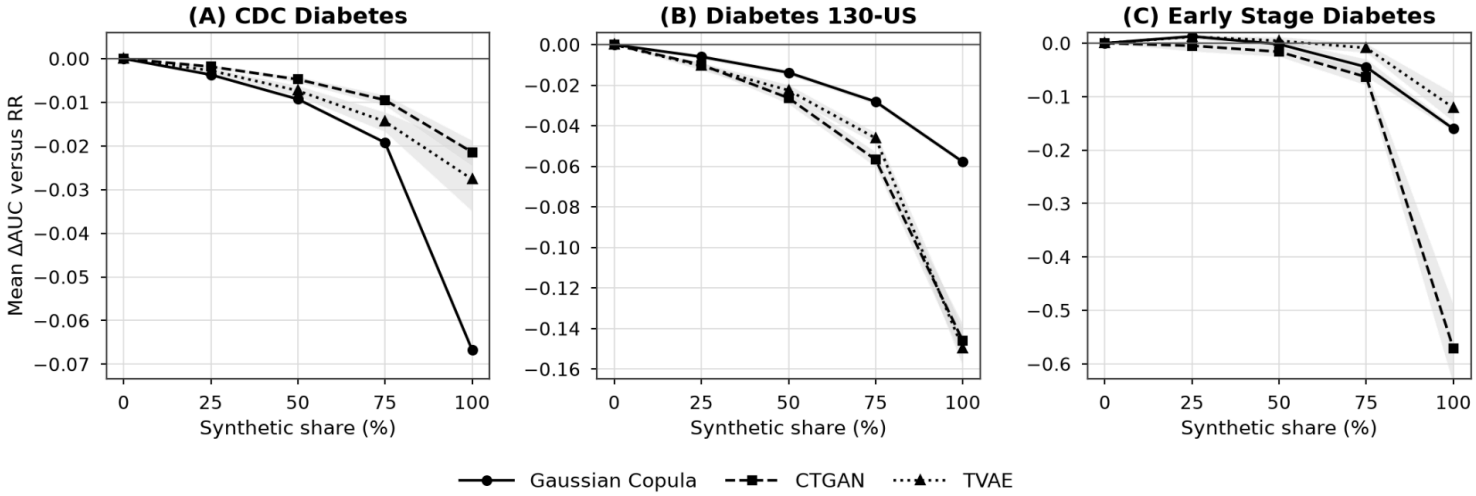


Figure 1

Mean ΔAUC relative to real-only training across fixed synthetic proportions. Lines identify generators using grayscale-safe styles and markers; bands are 95% bootstrap intervals across 10 replication-level means after averaging three downstream models. Vertical-axis ranges are dataset-specific.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM5ReproducibilityPackage.zip
- ESM1.pdf
- ESM4.xlsx
- ESM3.pdf
- ESM2.xlsx