Online Resource 3 — Supplementary figures

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

Journal: International Journal of Data Science and Analytics

Authors: Miguel Álvaro Andrade de Oliveira; Miguel Ricou; Alberto Freitas

Corresponding author: Miguel Álvaro Andrade de Oliveira; miguelalvaro13@gmail.com

Figures S1–S4 and their captions follow. The original scientific figure pages are preserved.

Online Resource 3 · Figure captions

Shared definitions. AUC denotes area under the receiver operating characteristic curve, evaluated on held-out real test data. The fixed-size classifier-training arms contain the same total number of rows, N: RR uses real data only; RS25, RS50 and RS75 replace approximately 25%, 50% and 75% of the real training rows with synthetic rows, respectively; SR uses synthetic data only. These are composition-controlled replacement comparisons, not increases in training-set size. S1–S3 are pooled descriptive summaries; their aggregation should not be interpreted as independent-dataset or replication-level inference. Error bars in S1 and S3 represent standard deviations, not confidence intervals (CIs), as specified in the archived plotting definitions.

Figure S1. Predictive utility by synthetic-data method and training composition

Mean test AUC for the real-only reference and the Gaussian Copula, CTGAN and TVAE training mixtures. Each bar pools the available dataset, classifier and replication records within its method–scenario combination; the RR baseline is displayed once under "reference". Error bars show one standard deviation of the pooled AUC values, reflecting dispersion across those records rather than uncertainty in an independently replicated overall treatment effect. Higher AUC indicates better discrimination.

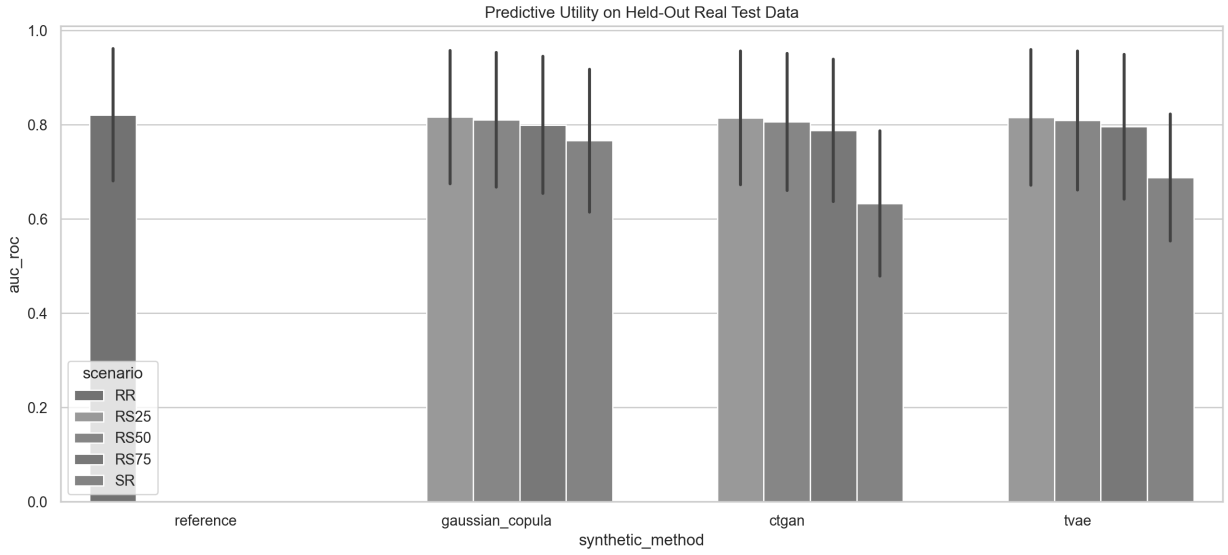


Figure S2. Distributional fidelity versus predictive utility

Each point represents a dataset–synthetic-method–scenario combination for RS25, RS50, RS75 or SR. The vertical coordinate is mean test AUC pooled across classifier and replication records within that combination. The horizontal coordinate is mean Hellinger distance averaged across the fidelity records for the corresponding dataset and synthetic method; this same fidelity summary is therefore used for all four training compositions. Lower Hellinger distance indicates closer distributional agreement, whereas higher AUC indicates better discrimination. The vertical axis is absolute AUC, not change from RR. Point styles identify training scenarios and the legend identifies synthesis methods. No confidence intervals are displayed; the relationship is descriptive, not a test of association.

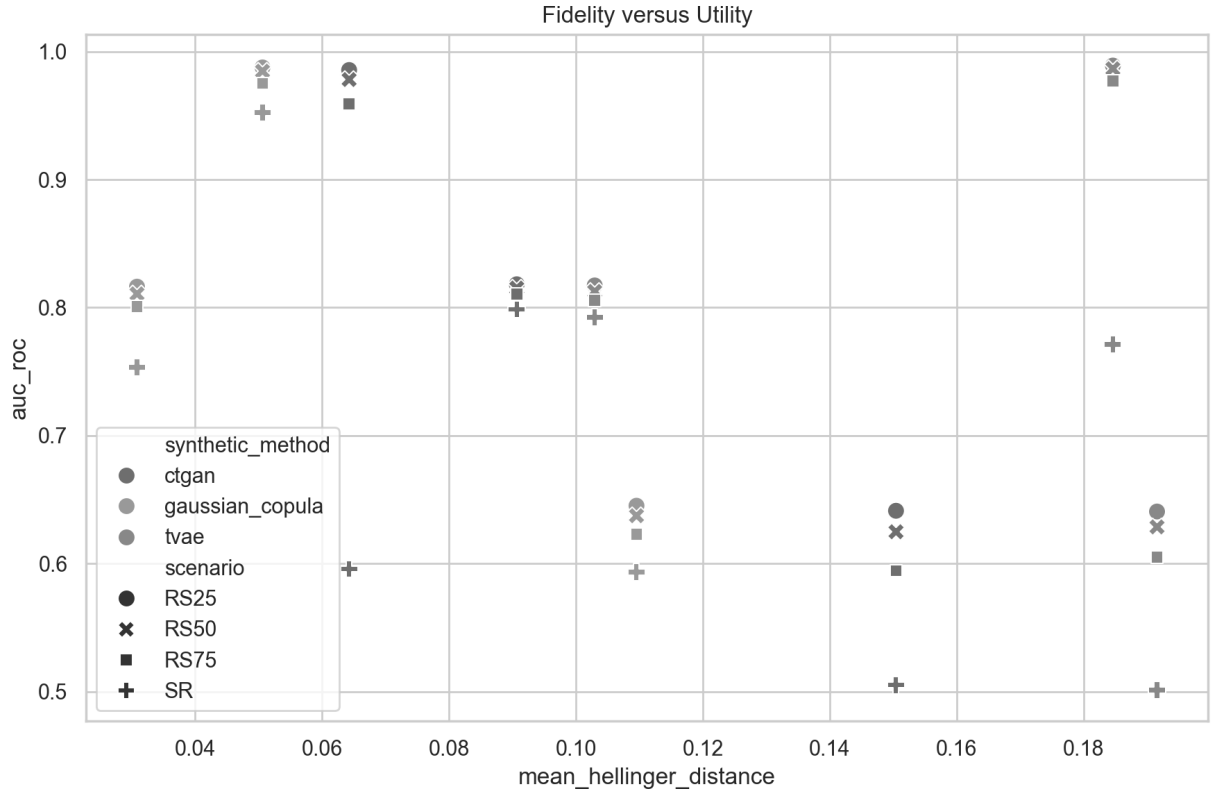


Figure S3. Within-subgroup change in AUC relative to real-only training

For each non-RR arm, subgroup AUC is compared with the RR value matched on dataset, classifier, replication, subgroup variable and subgroup value. Differences are calculated as AUC(non-RR) − AUC(RR), then averaged within each dataset–synthetic-method–scenario–subgroup-variable combination across the available classifiers, replications and subgroup values. Bars show the mean of these aggregated values for each displayed subgroup variable and scenario; error bars show one standard deviation across the contributing dataset–method summaries. Negative values indicate lower discrimination than RR within the evaluated subgroups. This is not a between-group disparity measure. Labels retain the source variable names and capitalization ("Age", "Sex", "age", "gender" and "race"); they should not be assumed to denote harmonized groups across datasets.

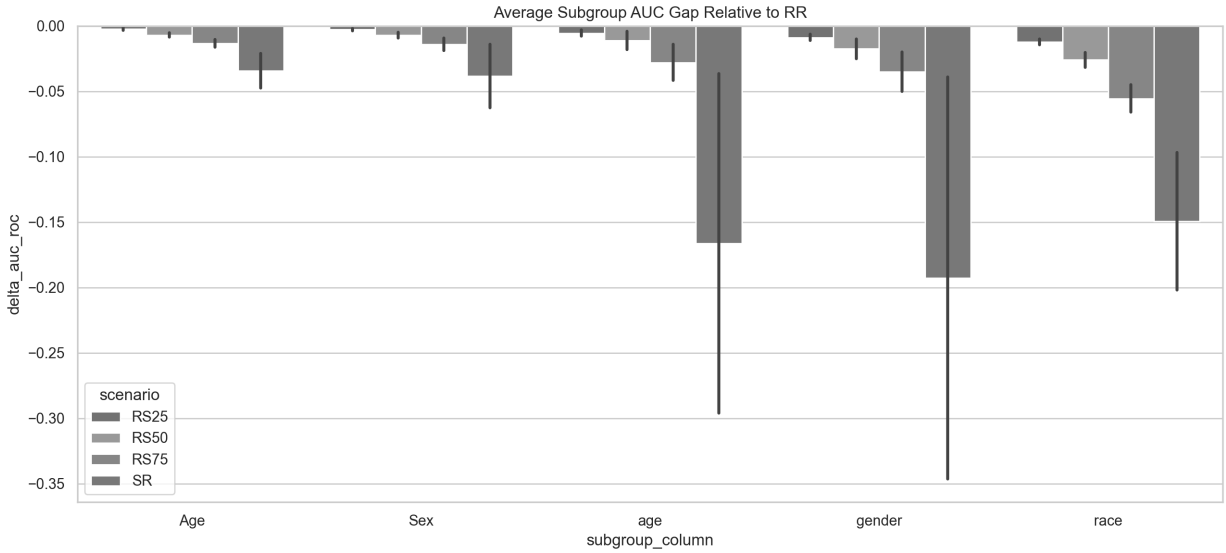


Figure S4. Fixed-size real–synthetic study design

(A) Schematic of the fixed real-data partition into training, validation and test sets (nominally 60%, 20% and 20%). Gaussian Copula, CTGAN or TVAE learns from the real training partition and provides synthetic rows for classifier training; validation and test data remain real. Three classifiers are evaluated on the real test set, with the total classifier-training size held at N. (B) RR, RS25, RS50, RS75 and SR vary the real–synthetic composition at that fixed size. Solid shading denotes real rows and hatching synthetic rows. The lower comparison illustrates the earlier RS (2N) design, which combined all N real rows with N synthetic rows and thus changed both composition and sample size. It is explanatory context, not an additional result in this figure. The schematic contains no empirical estimates or confidence intervals.

