

Supplementary information:

Collapse or creativity - Hallucinations in recursive synthetic data generation

Authors:

*Vibeke Binz Vallevik, University of Oslo/ DNV, Norway, Vibeke.Binz@gmail.com.

Aleksandar Babic, DNV, Norway.

Bjørn Edwin, Oslo University Hospital/University of Oslo.

Anne Kjersti Befring, University of Oslo, Department of Public law, Norway.

Jan F. Nygård, Norwegian Institute of Health/ University of Northern Norway, Norway.

1 Supplementary methods

1.1 Data Preprocessing and the population dataset

This appendix describes the full preprocessing workflow applied to raw MIMIC-III v1.4 data to construct a harmonized, ICU-stay-level dataset suitable for downstream modelling, to predict ICU readmission within 30 days using the earliest available measurements from each ICU stay.

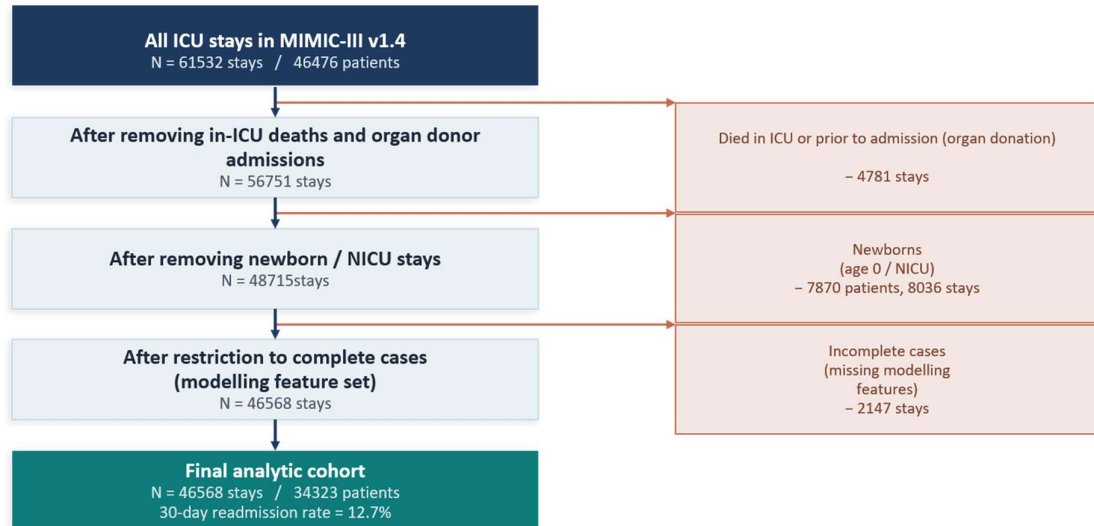
The steps were implemented in a modular Python pipeline integrates raw MIMIC tables, performs structured cleaning, applies physiologically informed outlier handling, and implements a transparent imputation strategy. This appendix summarizes the rationale and the key decisions, see code for further details[47].

1.1.1 Data source and cohort definition

We used the Medical Information Mart for Intensive Care III (MIMIC-III) public database, which contained deidentified clinical data of the patients who were admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts[22]. Access is credentialled and requires completion of human-subjects research training and acceptance of the PhysioNet data use agreement.

MIMIC-III contained hospital admission data from the period 2001 to 2012. A *patient* (subject) may have one or more *hospital admissions*; within an admission a patient may have one or more *ICU stays*. The unit of analysis in the final dataset is the ICU stay. We retained multiple ICU stays per patient, as each stay represents a distinct clinical episode with its own readmission risk.

We included all adult ICU stays (≥ 15 years). Newborns were excluded due to difference in physiology and measurements. Patients who died in the same hospital ICU were excluded, as these could not be readmitted even if measurements predict a severe status. Organ donor entries were also excluded (ICU entries that occur after registered time of death). Incomplete records were removed.



Supplementary figure 1 Cohort selection sequence

The source tables used were ADMISSIONS, PATIENTS, ICUSTAYS, CHARTEVENTS, LABEVENTS and DIAGNOSES_ICD - with D_ICD_DIAGNOSES supplying ICD-9 text descriptions. One manual correction was applied prior to ingestion: in the distributed ADMISSIONS file the free-text diagnosis field contains embedded delimiters that fragment the record across columns, and this field was reconstituted manually before processing.

1.1.2 Variables selection, extraction and harmonization

We extracted only the core variables as defined by Pilgram et. al¹³ for a lower complexity dataset (core dataset), and additional features with varying clinical relevance for a full dataset. We extracted demographics from *PATIENTS*, *ADMISSIONS*, and *ICUSTAYS* and first measurements (vital signs and laboratory values) from *CHARTEVENTS*, and *LABEVENTS*. All measurements carry single timestamps. For each variable, we selected the earliest valid timestamp in the ICU stay to reflect the patient's physiological state at ICU entry. If the earliest value was invalid according to the outlier boundaries, we used the next available value. Values were converted to comparable units. For height and weight transformations, certain assumptions were made regarding mis-registration rather than removing all invalid values. Weight was harmonized to kg: values recorded in pounds were converted using the standard factor, and implausibly large values were corrected for apparent hg coding. For height, the unit was inferred from the numeric value and transformed to cm.

Because measurement of vital values was recorded under multiple item identifiers in MIMIC-III, identifiers were grouped per vital concept and assigned an explicit priority order. For each ICU stay the earliest valid value was taken from the highest-priority identifier, with lower-priority identifiers used only to fill stays that would otherwise be missing (priority filling). See code for complete mapping of item identifiers to vitals. Measurements were restricted to those recorded

within the ICU admission–discharge window so that each value reflects the relevant care episode.

1.1.3 Quality control and physiologic validity bounds

Extracted numeric values were screened against pre-specified plausibility bounds defined per variable in the configuration file. Values outside the bounds were set to missing. These bounds are deliberately wide physiological limits intended to remove recording artefacts while allowing the wide variation typical of ICU patients, rather than to enforce clinically normal ranges. Bounds were defined through clinical expert consultation at Oslo University Hospital.

In addition to the outlier bounds we removed physiologically impossible values (e.g., negative labs, $\text{SpO}_2 > 100\%$, $\text{SBP} < \text{DBP}$). Diastolic and systolic blood pressure was resampled (pairwise) until $\text{DSP} < \text{SBP}$. Textual artifacts (“<5”, “>2000”, “error”) were set to missing. If the earliest measurement violated a physiologic bound, we selected the next timestamp. If all values for a variable were invalid, the variable was set to missing for that stay.

In the preprocessing pipeline, a maximum number of features from the core files was extracted to create a version of the dataset for quality control and sanity check of values, timestamps and patient stories prior to reducing to the analysis-ready dataset.

Supplementary table 1 Variables in dataset and outlier boundaries.

*The outlier bounds are extreme physiological bounds, not normal clinical range. *the resulting population dataset had haemoglobin values below 22, leading to an adjustment of reality limits for expert rule calculations.*

Variable name (“core” = used in both core and full datasets)	Categorical / Numerical	Description	Unit	Outlier bounds	Source
admission_type (core)	Cat	Type of hospital admission (e.g., emergency, elective).			PATIENTS
age (core)	Num	Patient age at admission.			PATIENTS, ADMISSIONS, ICUSTAY
albumin	Cat (ordinal)	Serum protein reflecting nutritional and inflammatory status.	g/dL	0.5 – 10	LABEVENTS
blood_urea_nitro (core)	Num	Marker of renal function and protein metabolism.	mg/dL	1 – 400	LABEVENTS
bmi	Cat (ordinal)	Body mass index derived from height and weight.	kg/m ²	5 – 150	Engineered feature, CHARTEVENTS
cholesterol (core)	Cat (ordinal)	Serum lipid level indicating metabolic and cardiovascular risk.	mg/dL	20 – 1000	LABEVENTS
creatinine (core)	Num	Renal function marker reflecting glomerular filtration.	mg/dL	0.1 – 30	LABEVENTS
diastolic_bp (core)	Num	Diastolic arterial blood pressure.	mmHg	10 – 230	CHARTEVENTS
ethnicity (core)	Cat	Ethnic background.			ADMISSIONS
first_careunit	Cat	Initial ICU location at admission.			ICUSTAY

Variable name ("core" = used in both core and full datasets)	Categorical / Numerical	Description	Unit	Outlier bounds	Source
gender	Cat (Boolean)	Male or female.			PATIENTS
glucose	Num	Blood glucose concentration indicating metabolic status.	mg/dL	5 – 2600	LABEVENTS
heartrate (core)	Num	Vitals measurement.	bpm	15 – 260	CHARTEVENTS
hemoglobin	Num	Oxygen-carrying capacity of blood.	g/dL	1.5 – 30*	LABEVENTS
icd9	Cat	Diagnostic code (on main chapter level) representing disease category for first diagnosis upon admittance.			Engineered. ADMISSIONS, DIAGNOSES_ICD, D_ICD_DIAGNOSES
nt-probnp (core)	Cat (ordinal)	Cardiac biomarker for heart failure severity.	pg/mL	0 – 100000	LABEVENTS
potassium (core)	Num	Serum potassium level reflecting electrolyte balance.	mEq/L	1.0 – 12	LABEVENTS
prior_icu	Cat (Boolean)	Indicator of previous ICU admissions.			Engineered.
readmission (core)	Cat, (Boolean)	ICU readmission within 30 days of ICU discharge.			Engineered.
respiratory_rate (core)	Num	Breaths per minute.	bpm	5 – 100	CHARTEVENTS
sodium	Num	Serum sodium level reflecting electrolyte and fluid balance.	mEq/L	90 – 200	LABEVENTS
spo2	Num	Peripheral oxygen saturation.	%	30 – 100	CHARTEVENTS
systolic_bp (core)	Num	Systolic arterial blood pressure.	mmHg	30 – 300	CHARTEVENTS

See frequency plots of the variables in the population dataset in supplemental data.

1.1.4 Missingness Handling and Imputation

To ensure reproducibility, the imputation strategy was built to ensure deterministic outputs. For patients with multiple ICU, cross-stay imputation was applied first - filling in missing values using the median of the same variable across other stays of the same patient to preserve patient-specific physiology.

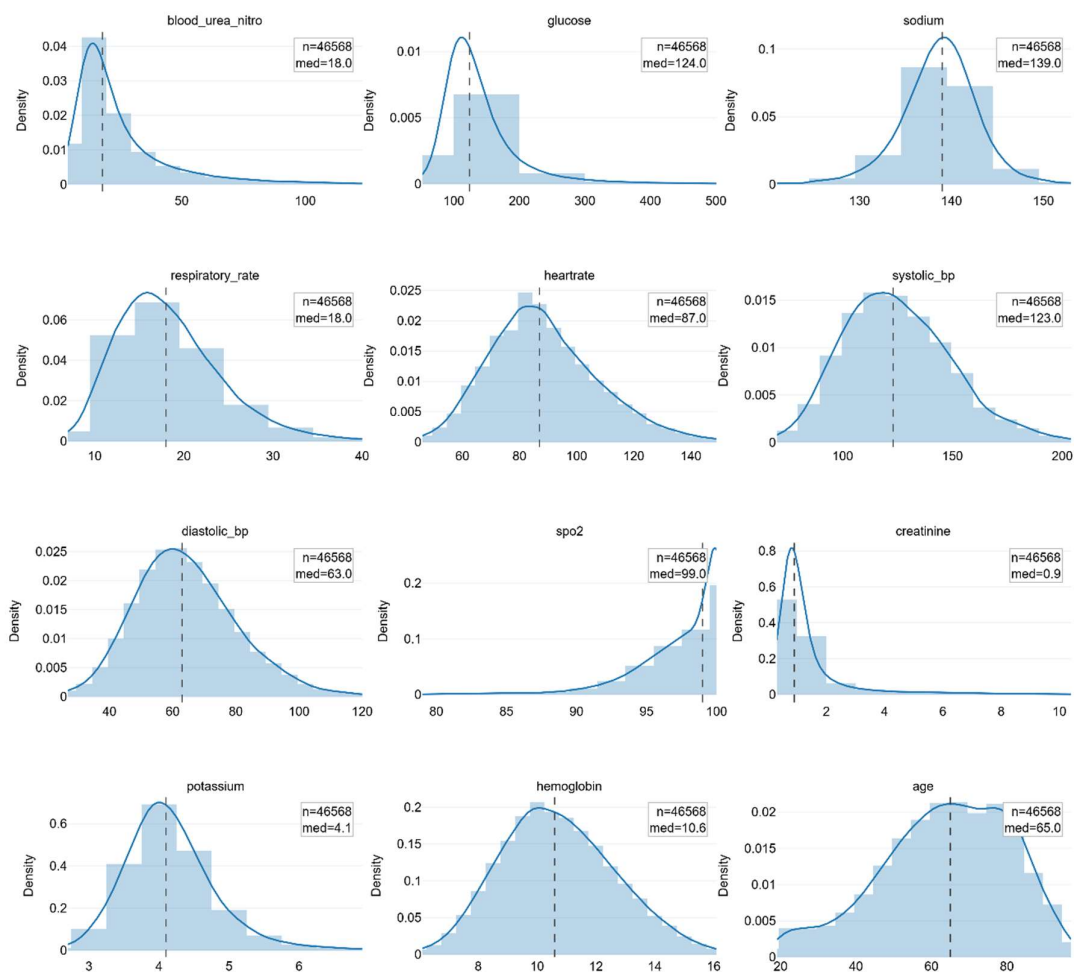
- Height: For missing height measurements, we applied gender specific synthetic value based on a narrow gaussian distribution to preserve patient-specific physiology but without major impact on BMI calculations. Female mean: 163cm, male mean 176 cm and corresponding standard deviations of 3.0 and 4.0.
- Weight: We used only real values from the same patient. No cross-patient imputation.
- Laboratory values and vital signs: We used deterministic median or nearest-value imputation depending on variable type.

- Categorization of missingness > 80 %: Features with larger than 80 % missingness were transformed to categorical features where missingness was a category, with the rationale that a missing test could have clinical significance: Albumin, Cholesterol and NT-proBNP. Clinician input was used to define meaningful classes and class boundaries, see code repository for details.
- Missingness: For the remaining features, the rows with missing data accounted for approximately 2 % of the dataset, and these rows were removed rather than imputed.

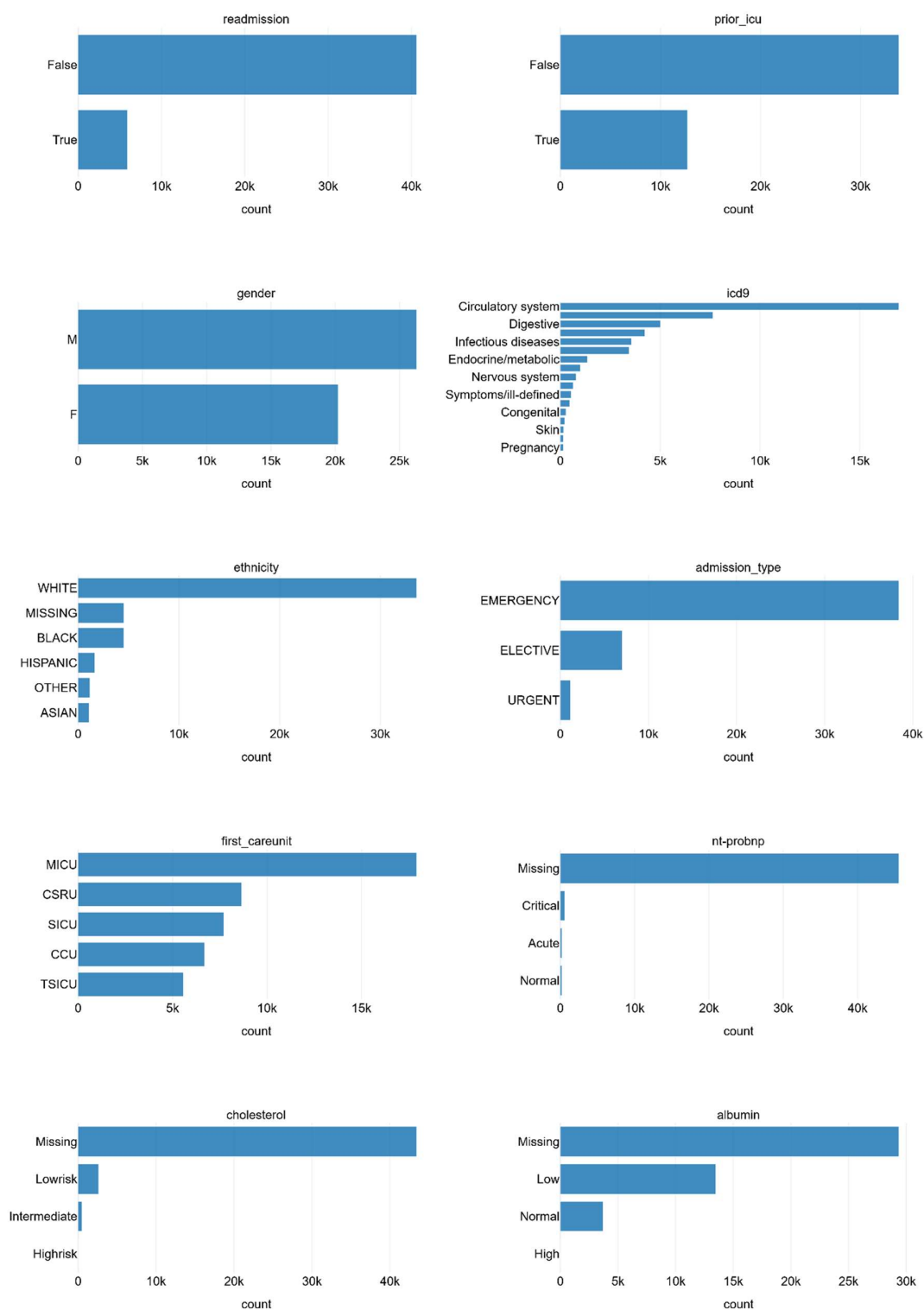
All imputation rules are implemented in the code repository.

1.1.5 Feature Engineering

- Age: Newborns were identified using `FIRST_CAREUNIT == "NICU"`. For adults < 89, age was calculated as `AGE = ADMITTIME_HADM.year - DOB.year`. MIMIC masks ages >89, so these patients were assigned a synthetic age drawn from a half-Gaussian distribution with mean 89 and a max age of 105. This was done only once for each patient, then calculated for consecutive ICU stays.
- BMI: BMI was calculated from weight and height. When a recorded weight was missing, weight classification was derived from ICD9 codes.
- ICD9: The first diagnosis was mapped to standard ICD9 groupings to reduce feature cardinality.
- Ethnicity: Ethnicity was grouped into larger groups to reduce feature cardinality.
- Readmission: A 30-day readmission flag was defined by ordering each patient's stays chronologically and identifying any subsequent ICU admission within 30 days of a prior ICU discharge. Organ-donor admissions, identified from the admission diagnosis text, were removed before this computation, since their returns are planned procedures rather than clinical readmissions.
- prior_icu: A flag was defined if the patient had previously been admitted to ICU.



Supplementary figure 2 Frequency plots for numerical variables in the population dataset.



Supplementary figure 3 Frequency plots for the categorical variables in the population dataset.

1.1.6 Reproducibility

All outputs can be regenerated from raw MIMIC-III. A fixed seed was used for all imputations to ensure reproducibility, details in code repository.

1.2 Clinical violation rules and LLM evaluator prompt

This appendix contains details related to the Expert rule script and LLM-based clinical evaluator. For full details, see the script[56].

Supplementary table 2 Clinical ranges for numeric variables.

"-" indicates that no threshold is defined in that direction for the variable. Normal ranges define standard reference intervals for each numeric variable, extreme thresholds mark values that are severely abnormal yet compatible with critical illness and fatal thresholds mark values at the margin of survivability. Reality limits define the absolute measurable physiological range for each variable; any value outside these limits is impossible regardless of context. BP = blood pressure.

Variable	Unit	Normal range	Extreme (low / high)	Fatal (low / high)	Reality limits
Heart rate	bpm	60–100	≤ 40 / ≥ 150	≤ 20 / ≥ 200	15–260
Systolic BP	mmHg	90–140	≤ 60 / ≥ 200	≤ 40 / —	30–300
Diastolic BP	mmHg	60–90	— / —	— / —	10–230
Respiratory rate	bpm	10–22	≤ 8 / ≥ 35	≤ 3 / ≥ 90	5–100
SpO ₂	%	95–100	≤ 80 / —	≤ 40 / —	30–100
Glucose	mg/dL	70–140	≤ 50 / ≥ 400	≤ 20 / ≥ 1500	5–2600
Potassium	mEq/L	3.5–5.0	≤ 2.5 / ≥ 7.5	— / ≥ 10	1.0–12
Sodium	mEq/L	136–145	≤ 120 / ≥ 155	≤ 100 / ≥ 180	90–200
Hemoglobin	g/dL	12.0–17.5	≤ 5.0 / ≥ 18.0	≤ 2.5 / —	1.5–25
Creatinine	mg/dL	0.6–1.2	— / ≥ 8.0	— / ≥ 20	0.1–30
Blood urea nitrogen	mg/dL	7–20	— / ≥ 80	— / ≥ 180	1–400
Age	years	18–89	— / —	— / —	0–120

Supplementary table 3 Overview of the rules that identify (a) impossible and b) suspicious combinations or values in the expert rule script.

In the dataset column, "Full only" indicates that the rule's conditions require at least one variable absent from the core datasets (SpO₂, haemoglobin, sodium, glucose, gender, ICD-9, BMI, albumin), so the rule cannot fire there. "Both" rules operate on both dataset variants, in some cases over a reduced variable set in core. Abbreviations: pp = pulse pressure (SBP – DBP); "normal" = within the normal reference range; "extreme"/"fatal" = beyond the extreme/fatal threshold for that variable; "non-extreme" = present and within the extreme thresholds.

Rule	Variables	Condition	Rule-name	Dataset
A) IMPOSSIBLE				
Reality limit	age; glucose; sodium; spo2; respiratory_rate; heartrate; systolic_bp; diastolic_bp; creatinine; blood_urea_nitro; potassium; hemoglobin	any value below/above its absolute physiological limit	reality_limit	Both ¹
Invalid categorical value	admission_type; first_careunit; ethnicity; icd9; bmi; nt-probnp; cholesterol; albumin; readmission; prior_icu; gender	categorical value not in the allowed set	bad_category	Both ¹
Diastolic ≥ systolic pressure	systolic_bp; diastolic_bp	dbp ≥ sbp	dbp_ge_sbp	Both
≤ 2 mmHg pulse pressure	systolic_bp; diastolic_bp; heartrate; respiratory_rate; spo2	pp ≤ 2 AND heartrate, respiratory_rate and spo2 all non-extreme	pp_zero	Full only
Impossible BUN/creatinine ratio	blood_urea_nitro; creatinine	bun/cr < 1.0	buncr_ratio_impossible	Both
Agonal respiration with normal saturation	respiratory_rate; spo2	rr ≤ 3 AND spo2 ≥ 98	rr_spo2_impossible	Full only
Fatal constellation with pristine vitals ²	heartrate; systolic_bp; respiratory_rate; spo2; glucose; potassium; sodium; hemoglobin; creatinine; blood_urea_nitro	≥ 3 fatal-range values, of which ≥ 1 is a primary vital, AND all primary vitals non-extreme	fatal_constellation_impossible	Both
Pregnancy in a male patient	icd9; gender	icd9 = Pregnancy AND gender = M	preg_male	Full only
Pregnancy outside reproductive age	icd9; age	icd9 = Pregnancy AND (age < 8 OR age > 60)	preg_age	Full only
B) SUSPICIOUS				
Low pulse pressure with normal vitals	systolic_bp; diastolic_bp; heartrate; respiratory_rate; spo2	pp < 5 AND heartrate, respiratory_rate, spo2 and systolic_bp all normal	pp_low	Full only
Wide pulse pressure in a younger adult	systolic_bp; diastolic_bp; age	pp > 150 AND sbp < 200 AND age < 60	pp_wide	Both
Hypertension with narrow pulse pressure	systolic_bp; diastolic_bp	sbp ≥ 200 AND pp < 10	pp_htn_narrow	Both
High BUN/creatinine ratio without anemia	blood_urea_nitro; creatinine; hemoglobin	bun/cr > 100 AND bun > 80 AND hb ≥ 11	buncr_ratio_high	Full only
Implausibly low creatinine	creatinine; age	cr ≤ 0.1 AND age ≥ 18	low_cr	Both
Uncompensated severe anemia	hemoglobin; heartrate; systolic_bp; respiratory_rate	hb ≤ 5 AND heartrate normal AND sbp < 90 AND rr ≤ 22	comp_anemia	Full only
Uncompensated shock	systolic_bp; heartrate; respiratory_rate	sbp < 50 AND heartrate normal AND rr ≤ 22	comp_shock	Both
Severe hyperkalemia with normal heart rate and renal function	potassium; heartrate; creatinine	k ≥ 7.5 AND heartrate normal AND cr < 1.0	comp_hyperk	Both

¹ Operates on both dataset variants over whichever variables are present; in the core datasets the check covers the reduced variable set (numeric: age, respiratory_rate, heartrate, systolic_bp, diastolic_bp, creatinine, blood_urea_nitro, potassium; categorical: admission_type, ethnicity, nt-probnp, cholesterol, readmission).

Rule	Variables	Condition	Rule-name	Dataset
Severe hyponatremia with shock and abnormal heart rate ³	sodium; systolic_bp; heartrate	na < 110 AND sbp < 90 AND (hr < 60 OR hr > 120)	comp_hypona	Full only
Cardiogenic collapse with preserved respiration	heartrate; systolic_bp; spo2; respiratory_rate	heartrate extreme low AND sbp extreme low AND spo2 non-extreme AND rr non-extreme	pair_cardiogenic	Full only
Hypertension with extreme bradycardia	systolic_bp; heartrate	sbp extreme high AND heartrate extreme low	pair_htn_brady	Both
Extreme tachycardia with extreme hypertension	heartrate; systolic_bp	heartrate extreme high AND sbp extreme high	pair_tachy_htn	Both
Fatal constellation with disturbed vitals	heartrate; systolic_bp; respiratory_rate; spo2; glucose; potassium; sodium; hemoglobin; creatinine; blood_urea_nitro	≥ 3 fatal-range values, of which ≥ 1 is a primary vital, AND ≥ 1 primary vital extreme	fatal_constellation_suspicious	Both
Multiple extreme values	heartrate; systolic_bp; respiratory_rate; spo2; glucose; potassium; sodium; hemoglobin; creatinine; blood_urea_nitro	≥ 5 extreme thresholds crossed	extreme_count	Both
Elderly female with very high hemoglobin	age; gender; hemoglobin	age > 60 AND gender = F AND hb ≥ 18	elderly_f_highhb	Full only
Pregnancy with high hemoglobin	icd9; hemoglobin	icd9 = Pregnancy AND hb > 15	preg_hb	Full only
Elective pregnancy admission	icd9; admission_type	icd9 = Pregnancy AND admission_type = ELECTIVE	pregnancy_elective	Full only
Unknown diagnosis with extreme laboratory values	icd9; nt-probnp; creatinine; potassium; hemoglobin	icd9 = UNKNOWN AND (nt-probnp = Critical OR cr extreme high OR hb extreme high OR hb extreme low)	unknown_extremelabs	Full only
Cardiac diagnosis without cardiac markers	icd9; cholesterol; nt-probnp; heartrate; systolic_bp	icd9 = Circulatory system AND cholesterol = Lowrisk AND nt-probnp = Normal AND heartrate normal AND sbp normal	circ_nomarkers	Full only
High-risk cholesterol with severe malnutrition	cholesterol; bmi; albumin; hemoglobin	cholesterol = Highrisk AND bmi = Underweight AND albumin = Low AND hb ≤ 7	nut_chol_maln	Full only
Obesity with hypotension and low albumin	bmi; systolic_bp; albumin; nt-probnp	bmi ∈ (Obese, Morbidly obese) AND sbp ≤ 90 AND albumin = Low AND nt-probnp = Normal	nut_obese_hypo	Full only
Low albumin with elevated BMI and severe anemia	albumin; bmi; hemoglobin; icd9	albumin = Low AND bmi ∈ (High, Obese, Morbidly obese) AND hb ≤ 5 AND icd9 ∉ protein-wasting chapters	alb_obese_anemia	Full only
High albumin in an underweight patient	albumin; bmi	albumin = High AND bmi = Underweight	alb_high_underwt	Full only
High albumin with a protein-wasting diagnosis	albumin; icd9	albumin = High AND icd9 ∈ (Neoplasms, Genitourinary, Digestive, Blood disorders)	alb_high_wasting	Full only
Severe hypokalemia with normal heart rate	potassium; heartrate	k ≤ 2.0 AND heartrate normal	severe_hypok	Both

Supplementary table 4 System prompt text for LLM evaluator script

Prompt

SYSTEM_PROMPT_FULL =

""You are a conservative clinical validation and hallucinations detection system evaluating SYNTHETIC patient records from an adult ICU cohort including critically ill and dying patients.

Numeric variable units: heartrate bpm, systolic_bp mmHg, diastolic_bp mmHg, respiratory_rate bpm, spo2 %, glucose mg/dL, potassium mEq/L, sodium mEq/L, hemoglobin g/dL, creatinine mg/dL, blood_urea_nitro mg/dL.

Categorical variables: readmission: 30-day readmission to ICU after this stay, prior_icu: previous ICU stay, gender: male(M) or Female (F), icd9 ICD-9-chapter, diagnosis group.

These rows have already passed an automated impossible-value filter. Flag ONLY rows with combinations or multivariate patterns that are physiologically impossible or clinically illogical in real human physiology, even in an ICU setting.

CRITICAL OUTPUT RULES:

- Output ONLY the numeric ID values
- Return them as a single comma-separated string
- No explanations, no reasoning, no bullet points, no other text
- If no rows are flagged, output exactly: NONE

Example correct output: 42,187,3901

"".strip()

1.3 Quality evaluation metrics

Evaluation metrics used in the study are documented in this section.

1.3.1 Frequency plots

Frequencies of synthetic versus real distributions were plotted per-feature and compared visually. For numerical features we used an overlay of KDE and bar chart. For categorical features: bar chart.

1.3.2 Coverage and precision density-profile comparison (α -Precision / β -Recall / Authenticity)

These metrics are a three-part framework that compares density profiles between real and synthetic data across quantile levels. The Synthcity⁴⁶ implementation follows the framework of Alaa et. al¹⁶. Three components make up this framework:

- α -precision: whether synthetic mass is well distributed relative to real data density profile; the metric penalizes synthetic mass over-concentrating in the innermost (core) region relative to real data (fidelity).
- β -recall: what fraction of real data density profile is covered by synthetic samples, whether entire regions of real-patient space are missing from the synthetic data (coverage, or "not missing real-world patients").
- Authenticity: a per-sample likelihood-ratio test estimating whether each synthetic record is a near-duplicate of one specific real training record, averaged across the synthetic dataset. This asks the question "is this a genuinely new patient, or a disguised copy of one specific real patient," closer in spirit to a privacy question.

All three range 0 to 1. Precision and recall are higher-is-better, meaning a closer density-profile match and more complete real-support coverage. Authenticity is also higher-is-better: higher means more genuinely novel generation and less memorization, while lower values mean synthetic samples sit suspiciously close to specific individual real records.

Comparing full density profiles across quantile levels, rather than a simple binary in/out-of-support test, catches distributional shape mismatches that PRDC's per-sample manifold test (below) can miss, the reason α -precision and β -recall sit under Distance alongside PRDC. Authenticity, asking a different question, memorization rather than distributional match, is why it's placed separately under Privacy, adding a memorization/privacy-adjacent signal present in neither PRDC nor the marginal-fidelity family above.

1.3.3 Downstream prediction utility -Train on Synthetic, Test on Real (TSTR-TRTR)

This measures whether a classifier trained entirely on synthetic data could substitute for one trained on real data, the most direct test of practical utility in this evaluation suite.

Hyperparameter optimization (HPO) pre-step (once per data seed $\times$ variant, shared across all 5 model seeds and all 20 generations within that data seed $\times$ variant): Bayesian hyperparameter optimisation via Optuna (50 trials, 5-fold inner cross-validation, TPE sampler) is run on the 10,000-row real training split, optimising jointly for mean CV AUROC. The best LightGBM⁴⁸ configuration is then evaluated once on the held-out 10,000-row real test set; this is the TRTR

baseline (Train on Real, Test on Real). The resulting parameters and decision threshold are saved for reuse.

Supplementary table 5 HPO search space (optimised jointly, objective = mean CV AUROC)

Parameter	Type	Range
boosting_type	categorical	gbdt, goss
num_leaves	int	4–60
max_depth	int	1–15
learning_rate	float (log)	2^{-8} –1.0
min_child_samples	int	1–60
min_data_in_leaf	int	1–60
reg_alpha (L1)	float	0–10
reg_lambda (L2)	float	0–10
feature_fraction	float	0.5–1.0
imbalance_method	categorical	none, scale_pos_weight, is_unbalance
early_stopping_rounds	int	7–30
use_target_encoding	bool	True / False
use_calibration (beta)	bool	True / False

TSTR, per generation: For each of the 20 generations, a new LightGBM is trained on that generation's 10,000 synthetic rows using the frozen best parameters and threshold from the HPO step. It is then evaluated on the same real test set. No re-optimisation occurs between generations. LightGBM was chosen to replicate the Pilgram et. al [17] study.

Why hyperparameters are frozen: re-running HPO at each generation would conflate data quality degradation with optimizer drift. Fixing the params means any TSTR decline reflects the synthetic data getting worse, not the tuner adapting to a collapsed distribution.

Metrics tracked, pooled across 15 chains per model type:

- Utility gap = TRTR AUROC – TSTR AUROC (↓ better; 0 = perfect substitution)
- TSTR AUROC, TSTR F1, and TSTR recall: absolute performance on the real test set
- TRTR AUROC shown as a dashed reference line, the real-data ceiling

This asks a practical, task-grounded question rather than a distributional one: if you trained the exact model you'd deploy on this synthetic data instead of the real data, how much predictive performance would you lose on the real task (readmission prediction)? Because the hyperparameters and threshold come from a single real-data HPO run and never change, any decline in TSTR across generations reflects the synthetic data's declining information content, not the model adapting differently to each generation.

The utility gap: lower is better, 0 means perfect substitution. TSTR's AUROC, F1, and recall: higher is better. Note that TRTR is not a hard ceiling at 1.0, it is itself well below 1.0 (a real, imperfect classifier on a real, imperfect task) and TSTR is not mathematically bounded by it either; a synthetic-trained classifier could in principle score above TRTR on the held-out real test set by chance, so "better than TRTR" is possible in the numbers even though it does not indicate the synthetic data is more informative than the real data.

TSTR/TRTR is the single most direct check in this evaluation suite of whether the degradation seen elsewhere, in marginal fidelity, coverage, or detection, costs anything on the downstream task the data exists to support. A generator could pass every distributional check and still fail

here, or vice versa. Limitation: this is evaluated against a single downstream task and model family (LightGBM, readmission prediction), so findings here may not generalize to other downstream tasks or model choices.

1.3.4 Marginal similarity - column-level distributional similarity

This metric provides an average score based on a per-column comparison of each synthetic column's distribution against the corresponding real column. Using the SDMetrics⁴⁹ implementations, it calculates the Kolmogorov-Smirnov (KS) complement for continuous columns and the Total Variation (TV) complement for discrete or categorical columns.

For continuous columns, KS complement reflects the maximum distance between the real and synthetic cumulative distribution functions, whether the overall shape and spread of values is preserved. For categorical columns, TV complement reflects the total variation distance between category frequencies, whether the relative prevalence of each category is preserved.

Both are normalized 0 to 1: 1.0 is a perfect match with identical distributions; 0.0 is maximum divergence with no distributional overlap.

This is a per-column diagnostic that directly identifies which specific variables are degrading, and how fast, across generations. The limitation is that these are 1D projections: a generator can preserve every column's individual distribution while completely abandoning the joint structure between columns.

1.3.5 Marginal similarity - bivariate structure

This metric aggregates the univariate KS-based column-shape similarity mentioned above with the column-pair trend similarity evaluated via SDMetrics [51] into a single quality score. Numerical-numerical pairs use a Spearman-correlation-based implementation; categorical-categorical and numerical-categorical pairs use a contingency-based similarity measure.

This tells us whether bivariate relationships between columns, which variables move together and how strongly, are preserved in the synthetic data, beyond each column's own isolated distribution.

This quality number ranges from 0 to 1; 1.0 indicates perfect preservation of both marginal shape and pairwise structure.

This captures a first layer of joint/multivariate structure that KS/TV complement alone cannot, though it's still limited to pairwise relationships and doesn't capture higher-order (3+ variable) joint structure.

1.3.6 Coverage and precision - nearest-neighbour manifold support (PRDC)

This is a four-part k-nearest-neighbour (k-NN) framework measuring fidelity and diversity via local neighbourhood support in the original feature space. The Synthcity implementation follows Naeem et. al [54]. Four components make up this framework:

- Precision: does each synthetic point fall inside at least one real point's k-NN neighbourhood? Fidelity: do individual synthetic records look like plausible real patients.
- Recall: does each real point fall inside at least one synthetic point's k-NN neighbourhood? Diversity and coverage: does the synthetic data collectively span the real data's range.

- Density: for each real point's k-NN neighbourhood, how many synthetic points fall inside it, summed and normalized by k. A more outlier-robust alternative to precision: a single unusual real point can inflate raw precision by creating an oversized neighbourhood that many synthetic points happen to fall into; density weights by how consistently supported that is.
- Coverage: the fraction of real points that have at least one synthetic point inside their own k-NN neighbourhood, using real-to-real distances only, not synthetic-to-real. A stricter, real-manifold-anchored alternative to recall.

Precision, recall, and coverage scores are bounded 0 to 1, higher is better. Density is not bound above by 1: a synthetic point sitting inside several overlapping real neighbourhoods is counted more than once, so values above 1 are expected and not an error.

Density and coverage directly correct for precision and recall's known sensitivity to outliers, giving a more robust fidelity/diversity read than the plain precision/recall pair. An important distinction: α -precision/ β -recall above and PRDC's precision/recall report different, non-interchangeable notions of "precision" and "recall." PRDC's is a permissive per-sample manifold-support test, α -precision/ β -recall's is a density-profile comparison, and both sit together under Distance, which is exactly why they should not be conflated despite the shared category and near-identical naming.

1.3.7 Aggregate distributional distance

This family complements the per-column comparisons of marginal fidelity and the local-neighbourhood comparisons of coverage and precision with a third view: how far apart the two full joint distributions sit, measured as a single number per generation rather than decomposed by feature or by neighbourhood. Three metrics contribute to this family, built on independent mathematical constructions rather than one shared method, so agreement or disagreement between them may serve as a form of triangulation rather than repetition.

Wasserstein distance

This metric evaluates the Sinkhorn-regularized Wasserstein-2 distance between the full real and synthetic patient cohorts. The calculation spans either 13 or 23 dimensions, depending on the variant, and is computed once per generation. We utilize an implementation that mirrors the Sinkhorn-OT algorithm from Synthcity [48], which has been specifically adapted to run without a GPU dependency.

This distance measures the minimal cost to reshape the synthetic point cloud into the real one, under squared Euclidean ground cost across all features jointly.

Lower is better; 0.0 indicates identical point clouds. Unlike the 0-1 bounded metrics used elsewhere in this study, Wasserstein distance is unbounded above and its absolute scale depends on the feature space's own units and dimensionality, so values are best read as a within-model, within-variant trend across generations rather than compared directly across variants with different feature counts.

Maximum Mean Discrepancy (MMD)

This is a kernel-based two-sample distance between the full real and synthetic cohorts, computed with a Gaussian radial basis function kernel. MMD [48] compares the two samples by

averaging kernel similarity within each sample and across the two, so it approaches zero as the distributions coincide and grows as they separate.

MMD measures the distance between the two cohorts' embeddings in a kernel feature space, computed from the full real-real, synthetic-synthetic, and real-synthetic kernel similarity matrices, each built from the complete feature vector per record rather than any single feature. All columns were standardised to zero mean and unit variance using the real reference only, and the kernel bandwidth was fixed by the median heuristic on that same standardised reference, then frozen across all generations, chains, and generators within each feature set.

Lower is better; 0.0 indicates identical distributions. Like Wasserstein distance, MMD is unbounded above and its absolute scale is not directly comparable across variants with different feature counts.

Presenting an optimal-transport-based distance and a kernel-based distance side by side, both confirmed to operate on the same joint feature space, provides two independently constructed estimates of the same underlying question, the joint-distribution half of the Distance category, alongside Wasserstein.

Jensen-Shannon divergence (JSD) – feature-averaged divergence

This metric captures the aggregate divergence between the real and synthetic cohorts. Using the Syndat⁵⁰ implementation, the Jensen-Shannon divergence is first evaluated for the real and synthetic distributions of each individual feature. Finally, these feature-level divergences are averaged to generate a single score per generation.

Because this metric is computed one feature at a time and then averaged, it captures how much each feature's own distribution has shifted, but not whether the relationships between features have changed. Two features could each individually match their real counterparts closely while their joint relationship diverges substantially, and this metric would not register that.

Higher indicates more divergence, the reverse of the 0-1 fidelity metrics used elsewhere in this study; 0.0 indicates identical distributions. The per-feature values that feed the average are not retained by the pipeline, so a feature-level breakdown of this metric is not available.

Positioned alongside Wasserstein distance and MMD, this metric isolate marginal-level divergence specifically, a quantity the other two members of this family do not separate out from joint-structure divergence on their own, the same marginal-versus-joint distinction. Distance category: this metric sits closer in spirit to KS/TV complement than to its own immediate neighbours, Wasserstein and MMD, despite all three sharing a row in the table.

1.3.8 Detection (real versus synthetic discriminability)

This evaluates the data via a classifier two-sample test. Four independent classifiers (a Gaussian Mixture Model, XGBoost, a multilayer perceptron, and logistic regression, implemented via Synthcity [48]. are evaluated using 5-fold stratified cross-validation. The maximum AUROC across the five folds is reported as the single value per classifier, per chain, and per generation. All four classifiers use fixed, untuned hyperparameters that remain identical across every model, chain, generation, and variant in the study.

Each classifier's AUROC measures whether it can learn a decision boundary separating real rows from synthetic rows. Because the four classifiers rely on different inductive biases (density-based, tree-based, neural, linear), they don't always agree; a separation caught by one but

missed by the others points to a specific, narrower structural difference rather than a wholesale loss of realism. The average detection score reported alongside the four classifiers compresses them into a single number, useful for at-a-glance tracking but able to mask divergent behaviour between them.

Higher AUROC means the classifier could tell real from synthetic more easily, so higher is worse (less realistic); 0.5 is the best case, indistinguishable from chance, and 1.0 is the worst case, trivially distinguishable.

Detection operationalizes "how similar are the two datasets" as "can a pattern-recognition system tell them apart, approached indirectly through classifiability rather than an explicit distance formula. It's a more holistic realism check than any single distributional-distance metric: a generator could preserve every marginal and pairwise trend checked elsewhere in this section and still be caught by a classifier picking up on higher-order structure. This is directly relevant to collapse and hallucination detection, since degraded synthetic output should become progressively easier for at least one classifier to flag.

1.3.9 Copies (NewRowSynthesis)

This metric (implemented via `SDMetrics`⁴⁹) evaluates the novelty of each synthetic row by checking for near-exact matches within the real data. A synthetic row is considered novel only if it does not simultaneously match any single real row across all fields. For a match to occur, categorical fields, boolean fields, and missing-value patterns must match exactly, while numerical fields are evaluated within a 1% tolerance. We compare the synthetic rows against their corresponding reference data, which remains fixed across all 20 generations.

The score is the fraction of synthetic rows that don't closely resemble any real row, a row-level check for genuinely new combinations versus near-copies of training data: a synthetic record that's a disguised copy of a specific real patient is a disclosure risk regardless of how statistically plausible it otherwise looks. It is unrelated to the hallucination-rate metric; the two are separate constructions, despite both speaking to novelty broadly.

Higher is better (more novel). Worth flagging here rather than in Results: the score sits at ceiling (1.0) in every cell checked, all models, chains, generations, both variants, no exceptions, because the joint condition of matching a real row within 1% tolerance across 13 to 23 columns at once becomes vanishingly unlikely regardless of generator. This is a property of the metric's construction interacting with table width, not a signal about model behaviour.

In principle, this is a useful complementary check to fidelity metrics: a generator that mode-collapses toward reproducing training rows would show this score dropping, a failure mode the distributional metrics elsewhere in this section don't directly target. In practice, given the ceiling saturation above, it carries no discriminating signal anywhere in this study.

1.3.10 Computational complexity

Alongside the quality and hallucination metrics, the pipeline logs resource-usage numbers for each training run: time spent training each generation's generator, and the time spent generating that generation's synthetic dataset.

These are resource-cost numbers, not quality signals: lower training or generation time reflects computational efficiency, not better synthetic data. They are included for transparency and

reproducibility, so a reader attempting to replicate or scale this experiment has a realistic sense of the computational cost involved.

1.4 Sensitivity analysis and statistical analysis

1.4.1 Time-to-utility-failure ranking

Ranking concordance across seed-pairs

We ranked the four generators for both the full and core dataset variants by time-to-failure within each of the 15 (data seed, model seed) pairs, and measured cross-seed-pair agreement with Kendall's coefficient of concordance (W). In the four-generator ranking, $W = 0.700$ ($\chi^2(3) = 31.5$, $p = 6.74 \times 10^{-7}$) for the full dataset and $W = 0.626$ ($\chi^2(3) = 28.2$, $p = 3.34 \times 10^{-6}$) for the core group: both rankings are more consistent across seed-pairs than chance, but this is driven overwhelmingly by RTVAE's near-universal first-ranked (fastest-to-fail) position (14/15 seed-pairs in both groups).

Because RTVAE's separation from the other three generators dominates the four-generator ranking, we repeated the analysis with RTVAE excluded, isolating whether ARF, CTGAN, and DDPM have a consistent internal order. Here the two dataset complexities diverge: $W = 0.287$ ($\chi^2(2) = 8.62$, $p = 0.0134$) in the full dataset, indicating a real but much weaker ranking agreement, versus $W = 0.0593$ ($\chi^2(2) = 1.78$, $p = 0.411$) in the core dataset, indistinguishable from a random ranking.

Independence of seed-pairs

The Friedman-equivalent test behind Kendall's W assumes each seed-pair is an independent rater. The 15 seed-pairs used here are not fully independent in that sense: they form a crossed design of 3 data seeds $\times$ 5 model seeds, so a seed-pair sharing a data seed or model seed with another is not an independent replicate of it. We quantified this by comparing pairwise rank agreement (Kendall's τ across all pairs of seed-pairs) for seed-pairs sharing a data seed, seed-pairs sharing a model seed, and seed-pairs sharing neither. Seed-pairs sharing a model seed showed measurably higher rank agreement than unrelated seed-pairs (gap of +0.199 for the four-generator core ranking and +0.419 for the three-generator core-excl-RTVAE ranking), confirming that the crossed design introduces correlation among the nominal 15 "raters." The reported Friedman/Kendall's W p -values should therefore be read as anticonservative - the seed-pair clustering inflates apparent significance to an extent we have not corrected for - rather than as p -values from 15 independent observations.

Survival-distribution comparison (log-rank and stratified Cox)

Kendall's W measures whether seed-pairs agree on a relative *order*; it does not test whether the underlying time-to-failure *distributions* differ. To test that directly, we ran a multivariate log-rank test (omnibus, with Holm-adjusted pairwise comparisons) and a Cox proportional-hazards likelihood-ratio test stratified on the (data seed, model seed) pair - stratifying on the seed-pair itself, rather than on data seed alone, restricts the comparison to within-seed-pair contrasts and so addresses the same non-independence identified above.

Excluding RTVAE (whose 100% failure rate produces complete separation in the Cox model), ARF, CTGAN, and DDPM's time-to-failure distributions differed significantly in the full dataset

(stratified Cox LRT: $\chi^2(2) = 8.72$, $p = 0.0128$; omnibus log-rank: $\chi^2(2) = 11.1$, $p = 0.00379$) but not in the core dataset ($\chi^2(2) = 2.03$, $p = 0.362$; log-rank $\chi^2(2) = 0.632$, $p = 0.729$). Holm-adjusted pairwise log-rank tests in the full dataset localized the difference to DDPM versus the other two (ARF vs DDPM: $p_{\text{holm}} = 0.00353$; CTGAN vs DDPM: $p_{\text{holm}} = 0.0399$), with ARF and CTGAN not distinguishable from each other ($p_{\text{holm}} = 0.268$); no pairwise comparison was significant in the core dataset.

Leave-one-data seed-out and leave-one-model seed-out sensitivity, see Supplementary table 19, showed the core-dataset null result was stable throughout (all $p > 0.10$ for both the log-rank and stratified Cox tests, across every excluded seed). The full dataset significant result was not fully robust: the stratified Cox LRT remained significant ($p < 0.03$) when excluding 3 of 5 model seeds individually, but rose to $p = 0.054$ and $p = 0.057$ for the other two, and rose above the conventional threshold when excluding data seed 1597 ($p = 0.0803$) or data seed 14930352 ($p = 0.137$), while remaining significant when excluding data seed 196418 ($p = 0.00441$). The same pattern held for Kendall's W on the same excl-RTVAE full subgroup (non-significant, $p = 0.0865$ and $p = 0.132$, when excluding those same two data seeds), see Supplementary table 7.

DDPM's rank

Within a seed-pair, a chain that never crosses the failure threshold is always ranked slower than one that does, regardless of how close it came to crossing. DDPM and DDPM_CORE's rank in the ranking tables therefore reflects each chain's failure/non-failure status, not an independently measured decline rate; see Supplementary table 6 for the corresponding failure counts and threshold-proximity behaviour referenced there.

Kendall's W as a coarse instrument

Kendall's W is computed over 3 items (the RTVAE-excluded groups) or 4 items (the full groups). With this few ranked items, W can only take a small number of distinct values, and W estimates from the 3-item and 4-item comparisons should not be read as directly comparable in magnitude.

Supplementary table 6 Per-generator rank summary by group.

Mean, median, minimum, and maximum rank (1 = fails fastest) across the 15 seed-pairs, and the number of seed-pairs in which each generator was ranked fastest-to-fail, for the full and core four-generator groups and for the corresponding three-generator groups excluding RTVAE/RTVAE_CORE.

* DDPM and DDPM_CORE's frequent slowest-to-fail rank follows from a comparatively low failure rate, not an independently measured slower decline - a chain that never crosses the failure threshold is always ranked slower than one that does (Methods), regardless of how close it came. See Table 1 for the corresponding threshold-proximity caveat.

Group	Generator	Mean rank	Median rank	Min rank	Max rank	N ranked fastest-to-fail	N seed-pairs
Full	ARF	2.566667	2.5	2	4	0	15
Full	CTGAN	2.866667	3	2	4	0	15
Full	DDPM	3.533333	4	1.5	4	0	15
Full	RTVAE	1.033333	1	1	1.5	14	15
Core	ARF_CORE	3.266667	3.5	2	4	0	15
Core	CTGAN_CORE	2.866667	3	2	4	0	15
Core	DDPM_CORE	2.833333	3	1.5	4	0	15

Group	Generator	Mean rank	Median rank	Min rank	Max rank	N ranked fastest-to-fail	N seed-pairs
Core	RTVAE_CORE	1.033333	1	1	1.5	14	15

Supplementary table 7 Kendall's coefficient of concordance (W) for generator failure-speed ranking, with leave-one-seed-out sensitivity.

W, the Friedman χ^2 statistic, degrees of freedom, and p-value for the full population and for each leave-one-data-seed-out and leave-one-model-seed-out subset, for all four groups (Full, core, Full excluding RTVAE, core excluding RTVAE).

* these p-values assume independent seed-pairs; see the SI text on seed-pair independence for a measured violation of that assumption.

Group	Subset	Kendall's W	N seed-pairs	N generators	Chi-square statistic	Degrees of freedom	p-value
Full	Full population	0.699537	15	4	31.47917	3	6.74E-07
Full	Excl. data seed 1597	0.662887	10	4	19.8866	3	0.000179
Full	Excl. data seed 196418	0.829167	10	4	24.875	3	1.64E-05
Full	Excl. data seed 14930352	0.657895	10	4	19.73684	3	0.000192
Full	Excl. model seed 987	0.711257	12	4	25.60526	3	1.15E-05
Full	Excl. model seed 28657	0.756466	12	4	27.23276	3	5.26E-06
Full	Excl. model seed 5702887	0.67971	12	4	24.46957	3	1.99E-05
Full	Excl. model seed 24157817	0.691304	12	4	24.88696	3	1.63E-05
Full	Excl. model seed 39088169	0.673851	12	4	24.25862	3	2.21E-05
Core	Full population	0.626107	15	4	28.17483	3	3.34E-06
Core	Excl. data seed 1597	0.654167	10	4	19.625	3	0.000203
Core	Excl. data seed 196418	0.647368	10	4	19.42105	3	0.000224
Core	Excl. data seed 14930352	0.607368	10	4	18.22105	3	0.000396
Core	Excl. model seed 987	0.654348	12	4	23.55652	3	3.09E-05
Core	Excl. model seed 28657	0.606932	12	4	21.84956	3	7.01E-05
Core	Excl. model seed 5702887	0.614307	12	4	22.11504	3	6.17E-05
Core	Excl. model seed 24157817	0.668116	12	4	24.05217	3	2.44E-05
Core	Excl. model seed 39088169	0.653017	12	4	23.50862	3	3.16E-05

1.4.2 Hallucination and utility disassociation

Across the 120 chains, most chains that failed by generation 19 never showed a qualifying hallucination spike (74/120, 61.7%), and a smaller subset showed a qualifying hallucination spike without meeting the failure criterion (17/120, 14.2%) (Supplementary Table 1). In the 16 chains where both events occurred, hallucination onset preceded failure in all 16 cases; it never

coincided with failure or followed it (Supplementary table 8). The median timing offset in these 16 chains was 7.5 generations (Wilcoxon signed-rank test on the paired timing offsets, $p = 0.0004$; Supplementary Table 2). Kendall's τ between hallucination-onset generation and failure generation across these 16 chains was 0.11 ($p = 0.59$; Supplementary Table 2). Leave-one-data seed-out and leave-one-model seed-out sensitivity checks left the Wilcoxon test significant ($p \leq 0.002$) and Kendall's τ non-significant ($p \geq 0.12$) in every subset (Supplementary table 9).

A Cox time-varying model with hallucination activity ($\text{Expert_any} \geq 0.05$) as a binary time-varying covariate, stratified by data seed and generator, was fit on 119 of 120 chains (89 failure events; one chain excluded for failing at generation 0, leaving no risk interval to model). Of the 119 chains, 33 contributed a within-observation transition in hallucination-active status. The hazard ratio for hallucination activity was 0.96 (95% CI 0.24–3.85, $p = 0.95$; Supplementary table 10).

Supplementary table 8: Per-chain dissociation between hallucination onset and utility failure.

Classification of all 120 chains (8 generator/variant tokens $\times$ 15 chains each: three data seeds $\times$ five model seeds) by whether hallucination onset and utility failure occurred, and, when both occurred, their relative generation order. Hallucination onset is the first single-generation increase in $\text{Expert_any} \geq 0.05$; utility failure is $\text{TSTR AUROC} \leq 0.5$ for two consecutive generations or complete loss of the readmission minority class.

Category	N chains
Both observed -- hallucination onset before failure	16
Both observed -- same generation	0
Both observed -- hallucination onset after failure	0
Failure only (no hallucination onset detected)	74
Hallucination onset only (no failure observed)	17
Neither event observed	13

Supplementary table 9 Paired timing of hallucination onset and utility failure.

Kendall's τ and paired Wilcoxon signed-rank test on hallucination-onset generation versus failure generation, restricted to chains where both events occurred ($n=16$ in the full population), with each data seed and each model seed excluded in turn. τ is computed on the two onset generations across chains; the Wilcoxon test is on the per-chain difference (failure generation – hallucination-onset generation).

Subset	n	Kendall's τ	τ p	Wilcoxon stat	Wilcoxon p
Full population	16	0.112	0.590376	0	0.000428
Excl. data seed 1597	10	0.422	0.121335	0	0.001953125
Excl. data seed 196418	11	0.069	0.793259	0	0.000976563
Excl. data seed 14930352	11	-0.021	0.933163	0	0.000976563
Excl. model seed 987	10	0.234	0.385742	0	0.001953
Excl. model seed 28657	12	-0.073	0.764476	0	0.000488
Excl. model seed 5702887	13	0.067	0.780719	0	0.000244
Excl. model seed 24157817	15	0.137	0.520803	0	0.000643
Excl. model seed 39088169	14	0.119	0.594293	0	0.000956

Supplementary table 10: Cox time-varying model of failure hazard under hallucination status.

Hazard ratio for utility failure with hallucination onset ($\text{Expert_any} \geq 0.05$) modelled as a binary time-varying covariate (0 before onset, 1 from onset onward), stratified by data seed and generator. Fit on 119 of 120 chains; one chain is excluded because it failed at generation 0, leaving no risk interval to model.

Covariate	Hazard ratio	95% CI (lower)	95% CI (upper)	p-value	Chains in model	Exclusion notes	Stratification
Hallucination active	0.9593	0.2388	3.8535	0.9532	119 of 120	1 chain(s) excluded (failure observed at generation 0 leaves no risk interval to model)	Data seed and generator

1.4.3 Hallucination and collapse coincidence

Across all 120 chains, onset was detected in 33 chains for hallucination rate (Expert_any , threshold 0.05, increase), 108 chains for beta-recall (threshold 0.0242, decrease), 71 chains for alpha-precision (threshold 0.0478, decrease), and 93 chains for authenticity (threshold 0.0254, increase) (Supplementary table 11). Hallucination onset was detected in DDPM (15/15 chains) for both full and core dataset variants, in one chain each in CTGAN and one for both full and core dataset RTVAE (33 chains total). No hallucination onset was detected in ARF, or the core dataset CTGAN chains (0/15 each) (Supplementary table 13).

Only DDPM with full or core dataset had enough chains with a defined hallucination onset to run the paired test (15/15 each); every other generator/variant token had 0 or 1 such chain, too few for a statistical test (**Supplementary table 13**). Onset detection requires an observed change from a previous generation, so the earliest generation at which onset can be detected for any metric is recursive generation 1. In DDPM, hallucination onset, beta-recall onset, and alpha-precision onset all occurred at generation 1 in every one of the 15 chains with no exceptions (Kendall's τ not defined, no variance in onset generation across chains; Wilcoxon test not applicable, every paired offset equal to 0). Authenticity onset in DDPM occurred at generation 1 in 6 of these 15 chains and at generation 2 in the remaining 9 (median offset 1 generation, Wilcoxon $p = 0.0027$; Kendall's τ not defined for the same reason). In the core dataset variant of DDPM, the generation of hallucination-onset varied across chains (range: generation 2–4). Kendall's τ was 0.47 ($p = 0.063$) for beta-recall, 0.80 ($p = 0.0012$) for alpha-precision, and 0.44 ($p = 0.064$) for authenticity. The paired Wilcoxon test was non-significant for beta-recall (median offset 0, $p = 0.180$) and for alpha-precision (median offset 0, $p = 0.157$), and significant for authenticity (median offset 1, $p = 0.0017$) (**Supplementary table 13**).

Pooled across all generator/variant tokens, restricted to chains with a defined onset in both metrics, Kendall's τ between hallucination-onset generation and QA-metric onset generation was 0.79 ($p < 0.001$, $n = 32$) for beta-recall, 0.96 ($p < 0.001$, $n = 31$) for alpha-precision, and 0.63 ($p < 0.001$, $n = 33$) for authenticity (**Supplementary table 12**). DDPM chains make up 30 of these 32 (beta-recall), 30 of 31 (alpha-precision), and 30 of 33 (authenticity) pooled chains; the remainder are single chains from CTGAN and RTVAE (**Supplementary table 13**). The median onset-generation offset in the pooled comparison was 0 generations for beta-recall and for alpha-precision, and 1 generation for authenticity, in the full population and in every leave-one-data seed-out and leave-one-model seed-out subset (Supplementary Table 5). The paired Wilcoxon test on this offset was non-significant in the full population for beta-recall ($p = 0.096$)

and for alpha-precision ($p = 0.157$), and significant for authenticity ($p < 0.001$). Kendall's τ remained significant ($p < 0.003$) for all three metrics in every leave-one-data seed-out and leave-one-model seed-out subset. The Wilcoxon test remained non-significant for alpha-precision in every subset ($p \geq 0.157$, or not computed in one subset where every paired offset was exactly 0) and significant for authenticity in every subset ($p < 0.001$); for beta-recall, the median offset remained 0 in every subset while the Wilcoxon p-value was non-significant in 6 of 8 leave-one-out subsets ($p \geq 0.096$) and significant in 2 (excluding data seed 1597, $p = 0.046$; excluding model seed 39088169, $p = 0.046$) (Supplementary Table 5).

Because the pooled comparison is restricted to chains with a defined hallucination onset, it does not include the majority of chains showing a beta-recall, alpha-precision, or authenticity onset without any detected hallucination onset (75, 40, and 60 chains respectively; Supplementary Table 4 minus Supplementary Table 6 totals).

Supplementary table 11 Onset crossing counts for hallucination rate and QA metrics.

Number of chains (of 120 total) with a detected onset for each metric. Onset is the first single-generation change, in that metric's calibrated direction, that meets its calibrated threshold: hallucination rate (Expert_any) increase ≥ 0.05 ; beta-recall decrease ≥ 0.0242 ; alpha-precision decrease ≥ 0.0478 ; authenticity increase ≥ 0.0254 .

Metric	Threshold	Direction of change at onset	N chains crossed	N chains total
Hallucination rate (Expert_any)	0.05	Increase	33	120
Beta-recall	0.0242	Decrease	108	120
Alpha-precision	0.0478	Decrease	71	120
Authenticity	0.0254	Increase	93	120

Supplementary table 12 Paired timing of hallucination onset and QA-metric onset, pooled across all generator/variant tokens.

Kendall's τ and paired Wilcoxon signed-rank test between hallucination-onset generation and each QA metric's onset generation, restricted to chains with a defined onset in both metrics, for the full population and with each data seed and each model seed excluded in turn. 30 of the pooled chains in every metric are DDPM chains with full or core dataset.

QA metric	Subset	N with both onsets defined	Kendall's tau	Kendall's tau p-value	Wilcoxon statistic	Wilcoxon p-value	Median offset (generations)
Beta-recall	Full population	32	0.791782	8.86E-07	3	0.095581	0
Alpha-precision	Full population	31	0.962838	4.00E-09	0	0.157299	0
Authenticity	Full population	33	0.630331	2.75E-05	0	1.21E-05	1
Beta-recall	Excluding data seed 1597	21	0.916693	6.74E-06	0	0.0455	0
Alpha-precision	Excluding data seed 1597	21	0.946997	3.31E-06	0	0.157299	0
Authenticity	Excluding data seed 1597	22	0.663525	0.000454	0	0.000132	1
Beta-recall	Excluding data seed 196418	21	0.785271	9.80E-05	1	0.654721	0

QA metric	Subset	N with both onsets defined	Kendall's tau	Kendall's tau p-value	Wilcoxon statistic	Wilcoxon p-value	Median offset (generations)
Alpha-precision	Excluding data seed 196418	21	1	7.73E-07			0
Authenticity	Excluding data seed 196418	22	0.682447	0.000297	0	0.000796	1
Beta-recall	Excluding data seed 14930352	22	0.720064	0.000214	3	0.095581	0
Alpha-precision	Excluding data seed 14930352	20	0.949027	2.93E-06	0	0.157299	0
Authenticity	Excluding data seed 14930352	22	0.565097	0.002081	0	0.000324	1
Beta-recall	Excluding model seed 987	25	0.754723	3.07E-05	2.5	0.157299	0
Alpha-precision	Excluding model seed 987	24	0.95265	3.33E-07	0	0.15719	0
Authenticity	Excluding model seed 987	25	0.674076	0.000112	0	0.0002	1
Beta-recall	Excluding model seed 28657	26	0.773272	1.42E-05	2.5	0.1573	0
Alpha-precision	Excluding model seed 28657	25	0.959391	1.05E-07	0	0.1573	0
Authenticity	Excluding model seed 28657	27	0.603541	0.000277	0	7.07E-05	1
Beta-recall	Excluding model seed 5702887	26	0.757143	2.92E-05	2.5	0.1573	0
Alpha-precision	Excluding model seed 5702887	25	0.976296	1.28E-07	0	0.3173	0
Authenticity	Excluding model seed 5702887	27	0.592101	0.00045	0	4.37E-05	1
Beta-recall	Excluding model seed 24157817	26	0.773272	1.42E-05	2.5	0.1573	0
Alpha-precision	Excluding model seed 24157817	25	0.976944	7.58E-08	0	0.3173	0
Authenticity	Excluding model seed 24157817	27	0.602492	0.000284	0	0.0002	1
Beta-recall	Excluding model seed 39088169	25	0.918492	9.83E-07	0	0.0455	0
Alpha-precision	Excluding model seed 39088169	25	0.951351	3.51E-07	0	0.1573	0
Authenticity	Excluding model seed 39088169	26	0.698089	4.48E-05	0	5.97E-05	1

Supplementary table 13 Per-generator breakdown of hallucination onset and QA-metric onset.
Paired comparison run separately for each of the 8 generator/variant tokens (≤ 15 chains each) instead of pooled.

Generator	QA metric	N tot	N (with hallucination onset)	N with QA-metric onset	N with both onsets defined	Kendall's tau	Kendall's tau p-value	Wilcoxon statistic	Wilcoxon p-value	Median offset (generations)
ARF	Beta-recall	15	0	15	0	n/a (n<2 chains with both onsets defined)				
ARF	Alpha-precision	15	0	15	0	n/a (n<2 chains with both onsets defined)				
ARF	Authenticity	15	0	15	0	n/a (n<2 chains with both onsets defined)				
ARF (CORE)	Beta-recall	15	0	15	0	n/a (n<2 chains with both onsets defined)				
ARF (CORE)	Alpha-precision	15	0	0	0	n/a (n<2 chains with both onsets defined)				
ARF (CORE)	Authenticity	15	0	9	0	n/a (n<2 chains with both onsets defined)				
CTGAN	Beta-recall	15	1	13	1	n/a (n<2 chains with both onsets defined)				
CTGAN	Alpha-precision	15	1	4	0	n/a (n<2 chains with both onsets defined)				
CTGAN	Authenticity	15	1	9	1	n/a (n<2 chains with both onsets defined)				
CTGAN (CORE)	Beta-recall	15	0	11	0	n/a (n<2 chains with both onsets defined)				
CTGAN (CORE)	Alpha-precision	15	0	4	0	n/a (n<2 chains with both onsets defined)				
CTGAN (CORE)	Authenticity	15	0	9	0	n/a (n<2 chains with both onsets defined)				
DDPM	Beta-recall	15	15	15	15	n/a (no variance in onset generation)		n/a (all offsets = 0 generations)	0	0
DDPM	Alpha-precision	15	15	15	15	n/a (no variance in onset generation)		n/a (all offsets = 0 generations)	0	0
DDPM	Authenticity	15	15	15	15	n/a (no variance in onset generation)		0	0.0027	1
DDPM (CORE)	Beta-recall	15	15	15	15	0.4678	0.0629	3	0.1797	0
DDPM (CORE)	Alpha-precision	15	15	15	15	0.8036	0.0012	0	0.1573	0
DDPM (CORE)	Authenticity	15	15	15	15	0.4416	0.0637	0	0.0017	1
RTVAE	Beta-recall	15	1	14	1	n/a (n<2 chains with both onsets defined)				
RTVAE	Alpha-precision	15	1	13	0	n/a (n<2 chains with both onsets defined)				
RTVAE	Authenticity	15	1	15	1	n/a (n<2 chains with both onsets defined)				
RTVAE (CORE)	Beta-recall	15	1	10	0	n/a (n<2 chains with both onsets defined)				
RTVAE (CORE)	Alpha-precision	15	1	5	1	n/a (n<2 chains with both onsets defined)				
RTVAE (CORE)	Authenticity	15	1	6	1	n/a (n<2 chains with both onsets defined)				

1.4.4 Time between samples

First-measured values were extracted independently for each variable. Because laboratory and vital signs are not sampled simultaneously in ICU workflows, these values may represent

different physiologic time points. Some biomarkers change rapidly (minutes-hours), while others evolve over longer intervals (hours-days). The following variable timestamps were considered time sensitive and measured: Blood pressure (BP), Heartrate, Respiratory rate, SpO2, Potassium, Sodium, Glucose, Creatinine, Blood Urea Nitrogen, Haemoglobin. For each row, we measured the variance in time stamps, see **Error! Reference source not found.**

Supplementary table 14 Row-level spread between relevant timestamps. (hours)

condition	mean_hr	median_hr	std_hr	var_hr2	min_hr	max_hr	p90_hr	p95_hr
with_sbp_dbp	12.622	7.7	25.557	653.153	0	1012.5	23.5	38.5
without_sbp_dbp	8.293	6.467	10.218	104.405	0	697.05	16.65	21.483

We found the biggest discrepancy with blood pressure, as the pairwise draw requirement means the first measured value was sometimes not eligible. The table shows result including and excluding BP. A subset of expert rules was sensitive to timestamp asynchronies between SBP/DBP and other vitals. Rules requiring synchronous hypotension and bradycardia or shock compensation (e.g., comp_shock, comp_anemia, pair_cardiogenic) showed the highest risk. Rules requiring multiple extreme values (e.g., fatal constellations) were more robust, as other variables compensated for potential SBP/DBP drift.

The rules listen in **Error! Reference source not found.** include two or more of the time critical variables, and may therefore have increased uncertainty due to sampling times of labs and vitals. The mean count of times the rule has fired pr dataset of 10,000 records is also listed.

Supplementary table 15 List of the rules that fired violations in the expert rule that include more than two time-critical variables.

These are rules that may have been affected by time gap between measured values. The count of fired rules is reported in mean value rounded to 0 decimals per dataset of 10,000 records. Several rules can fire pr record, so a count of 1000 does not necessarily add up to 10 %. HB = haemoglobin, BUN = Blood urea nitrogen, HR = heartrate. Resp = respiratory rate.

Rule	Time sensitive variables in rule	Number of times rule fired mean full dataset (mean core dataset)			
		ARF (ARF_CORE)	CTGAN (CTGAN_CORE)	DDPM (DDPM_CORE)	RTVAE (RTVAE_CORE)
alb_obese_anemia (suspicious): Low albumin with elevated BMI and severe anemia	albumin; HB	5		843	6
buncr_ratio_high (suspicious): High BUN/creatinine ratio without anemia	BUN; Creatinine; HB	8	56	788	2
buncr_ratio_impossible (impossible): Impossible BUN/creatinine ratio	BUN; Creatinine	36(29)	58(22)	1899(869)	14(68)
circ_nomarkers (suspicious): Cardiac diagnosis without cardiac markers	HR; SBP	0	5	0	
comp_anemia (suspicious): Uncompensated severe anemia	HB; HR; SBP; resp	10	0	13	1
comp_hyperk (suspicious): Severe hyperkalemia with normal heart rate and renal function	potassium; HR; Creatinine	1(1)	1(0)	28(214)	1(0)
comp_hypona (suspicious): Severe hyponatremia with shock and abnormal heart rate	sodium; SBP; HR		0	2318	
comp_shock (suspicious): Uncompensated shock	SBP; HR; resp	0(0)		36(188)	0(33)

Rule	Time sensitive variables in rule	Number of times rule fired mean full dataset (mean core dataset)			
		ARF (ARF_CORE)	CTGAN (CTGAN_CORE)	DDPM (DDPM_CORE)	RTVAE (RTVAE_CORE)
extreme_count (suspicious): Multiple extreme values	HR; SBP; resp; spo2; glucose; potassium; sodium; HB; Creatinine; BUN	0		8563(2222)	0(0)
fatal_constellation_suspicious: Fatal constellation with disturbed vitals	HR; SBP; resp; spo2; glucose; potassium; sodium; HB; Creatinine; BUN			6039(1190)	
nut_chol_maln (suspicious): High-risk cholesterol with severe malnutrition	cholesterol; albumin; HB			0	
nut_obese_hypo (suspicious): Obesity with hypotension and low albumin	SBP; albumin; NT-proBNP	0	1	35	
pair_cardiogenic (suspicious): Cardiogenic collapse with preserved respiration	HR; SBP; spo2; resp	0		13	
pair_htn_brady (suspicious): Hypertension with extreme bradycardia	SBP; HR	0(0)		2220(1295)	0
pair_tachy_htn (suspicious): Extreme tachycardia with extreme hypertension	HR; SBP	0(0)	0	2308(1386)	1(1)
pp_low (suspicious): Low pulse pressure with normal vitals	SBP; DBP; HR; resp; spo2	3	10	0	0
pp_zero (impossible): Zero pulse pressure	SBP; DBP; HR; resp; spo2	42	29	6	4
severe_hypok (suspicious): Severe hypokalemia with normal heart rate	Potassium; HR	9(3)	(0)	62(511)	(14)

1.5 Calibrations

1.5.1 Calibration of the expert rules script on real data

When running the expert rules script on the three dataset partitions, data seed 14930352 flagged 15 records as suspicious, data seed 1597 flagged 14 records as suspicious and dataset 196418 flagged 24 records as suspicious. The rules fired are disclosed in Supplementary table 16. Rule definitions are described in SI section 1.2.

Supplementary table 16 Clinical implausibility rules that were flagged for the three real datasets.

Rule fired on real data	Mean ($\pm$ 95% CI)
alb_high_underwt	0.33 ($\pm$ 1.43)
alb_high_wasting	0.67 ($\pm$ 0)
alb_obese_anemia	2.33 ($\pm$ 0)
circ_nomarkers	0.67 ($\pm$ 0)
comp_anemia	0.67 ($\pm$ 0)
comp_hyperk	2.33 ($\pm$ 0)
comp_shock	0.33 ($\pm$ 0)
extreme_count	0.33 ($\pm$ 0)
low_cr	6.67 ($\pm$ 0)
nut_obese_hypo	0.67 ($\pm$ 0)

pair_htn_brady	0.67 ($\pm$ 0)
pair_tachy_htn	0.67 ($\pm$ 0)
pp_wide	0.33 ($\pm$ 0)
preg_hb	0.33 ($\pm$ 0)
severe_hypok	1.33 ($\pm$ 0)

1.5.2 Calibration of the LLM evaluator

The LLM evaluator was used only for illustration purposes due to its high cost. It was used extensively while developing the scripts and in preprocessing of the population data. On the final version of the real dataset an observational test was run for documenting calibration on data partition from data seed 14930352. Of 10,000 records the evaluator identified 304 records as suspicious in addition to the 15 filtered expert-rule identified records.

To investigate the overlap between hallucination identified by expert rules and LLM evaluator, we ran a pure LLM evaluation without screening for expert rules already identified, and compared to the expert-rule identifications. As a calibration case, we used CTGAN generation 19 which showed relatively stable results. Three different data seed-model seed combinations gave a mean of 41.09% [95 % CI 28.83, 53.35] expert script flagged rows that were also flagged by the LLM, and 40.24% [26.02, 54.46] LLM rows also flagged by rules.

1.5.3 Overlap of rules identified by HR and expert rules

Supplementary table 17 quantifies how well hallucination rates measured by HR (bin=5) covers records flagged as clinically implausible by the expert rules. Rule definitions are detailed in section 1.2. The results for full and core dataset are reported separately, as the core dataset has a more limited number of rules. Pooled across chains, HR (bin=5) missed 0.13 % of the expert-rule flagged records in the full dataset, and 11.93% in the core dataset. For generator-specific results, see Supplementary table 18.

Supplementary table 17 overlap between records identified as hallucinations between HR (bin=5) and expert rules.

Category is the rule's severity bucket (impossible vs. suspicious). n chains is the number of chains (out of up to 60 per dataset -15 chains $\times$ 4 generators) in which the rule fired at least once. Overlap ratio is the fraction of that rule's flagged records also flagged by HR (bin=5); Miss rate is its complement (1 – overlap ratio). Both are computed at the individual-record level, pooled across every chain and generation in which the rule fired. The 95% CI on miss rate is a chain-level bootstrap (chains resampled with replacement, 2000 resamples); a zero-width CI reflects perfect consistency across chains, not a computation error. Rows are sorted from highest to lowest miss rate within each dataset.

Datas et	category	Rule	n chains	Overlap ratio	Miss rate (mean)	Miss rate mean – 95 % CI	Miss rate mean + 95 % CI
core	suspicious	comp_shock	20	0.7074	0.2926	0.2134	0.4173
core	suspicious	low_cr	38	0.7846	0.2154	0.0988	0.329
core	suspicious	severe_hypok	34	0.8249	0.1751	0.1478	0.2194
core	suspicious	pp_wide	38	0.8257	0.1743	0.0923	0.1862
core	impossible	dbp_ge_sbp	50	0.8598	0.1402	0.0924	0.1968
core	impossible	buncr_ratio_impos sible	50	0.9195	0.0805	0.0197	0.2041
core	suspicious	comp_hyperk	44	0.9697	0.0303	0.0267	0.0373
core	suspicious	pair_htn_brady	30	0.9999	0.0001	0.0001	0.0002
core	suspicious	pair_tachy_htn	28	1	0	0	0.0001

Datas et	category	Rule	n chains	Overlap ratio	Miss rate (mean)	Miss rate mean – 95 % CI	Miss rate mean + 95 % CI
core	suspicious	extreme_count	17	1	0	0	0
core	suspicious	fatal_constellation _suspicious	15	1	0	0	0
full	impossible	pp_zero	54	0.9926	0.0074	0.0057	0.009
full	suspicious	low_cr	45	0.9954	0.0046	0.0017	0.0096
full	suspicious	pp_low	48	0.9961	0.0039	0.0021	0.0059
full	suspicious	comp_shock	16	0.9968	0.0032	0.0013	0.0055
full	suspicious	severe_hypok	28	0.997	0.003	0.0022	0.0039
full	impossible	buncr_ratio_impos sible	48	0.9989	0.0011	0.0006	0.0024
full	impossible	dbp_ge_sbp	51	0.9995	0.0005	0.0003	0.0008
full	suspicious	comp_anemia	33	0.9996	0.0004	0	0.0015
full	suspicious	pp_wide	46	0.9997	0.0003	0	0.001
full	suspicious	buncr_ratio_high	53	0.9999	0.0001	0	0.0002
full	suspicious	alb_obese_anemia	34	1	0	0	0
full	impossible	preg_age	40	1	0	0	0
full	impossible	preg_male	41	1	0	0	0
full	suspicious	alb_high_underwt	23	1	0	0	0
full	suspicious	alb_high_wasting	30	1	0	0	0
full	suspicious	circ_nomarkers	34	1	0	0	0
full	suspicious	comp_hyperk	44	1	0	0	0
full	suspicious	comp_hypona	16	1	0	0	0
full	suspicious	elderly_f_highhb	29	1	0	0	0
full	suspicious	extreme_count	19	1	0	0	0
full	suspicious	fatal_constellation _suspicious	15	1	0	0	0
full	suspicious	nut_chol_maln	1	1	0	0	0
full	suspicious	nut_obese_hypo	36	1	0	0	0
full	suspicious	pair_cardiogenic	30	1	0	0	0
full	suspicious	pair_htn_brady	31	1	0	0	0
full	suspicious	pair_tachy_htn	32	1	0	0	0
full	suspicious	preg_hb	29	1	0	0	0
full	suspicious	pregnancy_electiv e	38	1	0	0	0

Supplementary table 18Supplementary table 17 provides an additional per-generator breakdown of the same record-level miss rate reported in **Supplementary table 17** - the percentage of a given rule's flagged records that HR (bin=5) failed to also flag, pooled across that generator's chains. Missed records rate is a share of that rule's own flagged records, not a share of the generator's overall hallucination rate. Only rules with at least one flagged record for that generator are listed; rows are sorted from highest to lowest missed-records rate within each generator. Full and core variants of the same base generator are listed as separate rows, reflecting their different feature sets.

Supplementary table 18 per-generator overview of expert rules that are missed by HR bin=5.
Missed records rate is the % of the expert rules that were missed, not a total hallucination percentage.

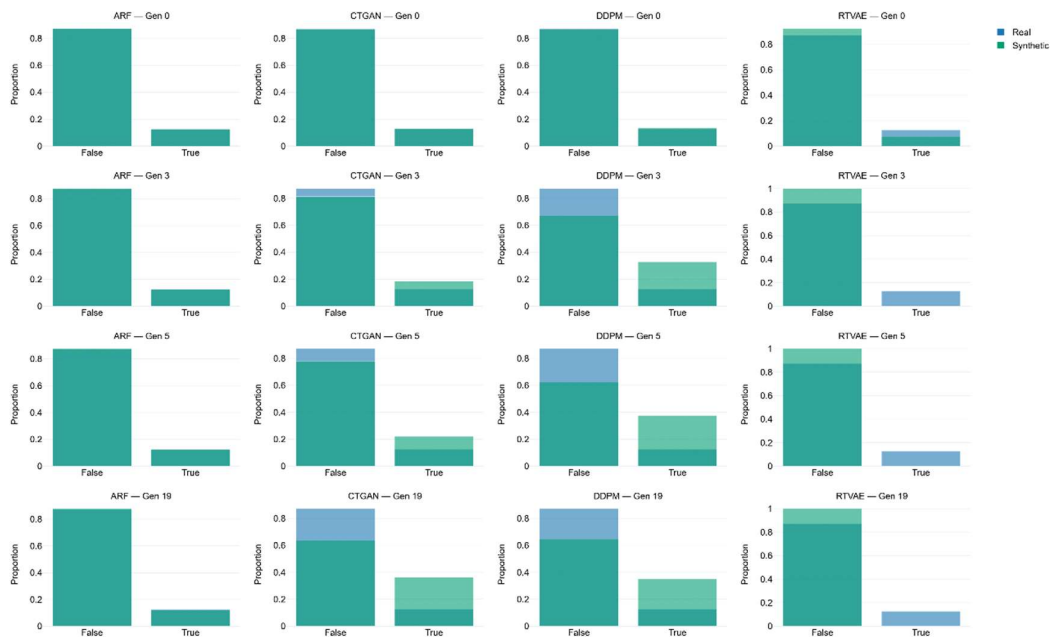
generator	category	rule	Missed records rate (mean)
ARF	suspicious	low_cr	0.0488
ARF	impossible	buncr_ratio_impossible	0.0386
ARF	suspicious	comp_shock	0.0141
ARF	impossible	pp_zero	0.0096
ARF	suspicious	pp_low	0.007
ARF	suspicious	buncr_ratio_high	0.004
ARF	suspicious	severe_hypok	0.0037
ARF	impossible	dbp_ge_sbp	0.0029
ARF	suspicious	alb_obese_anemia	0.0007
ARF	suspicious	pp_wide	0.0004
ARF (CORE)	suspicious	low_cr	0.833
ARF (CORE)	impossible	buncr_ratio_impossible	0.7378
ARF (CORE)	suspicious	comp_shock	0.5882
ARF (CORE)	suspicious	severe_hypok	0.5538
ARF (CORE)	suspicious	comp_hyperk	0.4648
ARF (CORE)	impossible	dbp_ge_sbp	0.402
ARF (CORE)	suspicious	pair_tachy_htn	0.1765
ARF (CORE)	suspicious	pair_htn_brady	0.1293
ARF (CORE)	suspicious	pp_wide	0.1062
CTGAN	suspicious	low_cr	0.0191
CTGAN	impossible	buncr_ratio_impossible	0.0147
CTGAN	impossible	pp_zero	0.0046
CTGAN	suspicious	pp_low	0.003
CTGAN	impossible	dbp_ge_sbp	0.0016
CTGAN	suspicious	buncr_ratio_high	0.0007
CTGAN (CORE)	suspicious	severe_hypok	0.8
CTGAN (CORE)	suspicious	low_cr	0.7996
CTGAN (CORE)	impossible	dbp_ge_sbp	0.4342
CTGAN (CORE)	suspicious	comp_hyperk	0.3966
CTGAN (CORE)	impossible	buncr_ratio_impossible	0.2628
CTGAN (CORE)	suspicious	pp_wide	0.1344
DDPM	suspicious	comp_shock	0.0029
DDPM	suspicious	severe_hypok	0.0029
DDPM	suspicious	low_cr	0.001
DDPM	suspicious	comp_anemia	0.0005
DDPM (CORE)	suspicious	comp_shock	0.2359
DDPM (CORE)	suspicious	severe_hypok	0.1626
DDPM (CORE)	suspicious	low_cr	0.0933
DDPM (CORE)	suspicious	pp_wide	0.0612
DDPM (CORE)	impossible	dbp_ge_sbp	0.0302
DDPM (CORE)	suspicious	comp_hyperk	0.0295
DDPM (CORE)	impossible	buncr_ratio_impossible	0.0013
DDPM (CORE)	suspicious	pair_htn_brady	0.0001
RTVAE	impossible	pp_zero	0.0066

generator	category	rule	Missed records rate (mean)
RTVAE	impossible	buncr_ratio_impossible	0.0034
RTVAE	impossible	dbp_ge_sbp	0.0032
RTVAE (CORE)	impossible	buncr_ratio_impossible	0.7513
RTVAE (CORE)	suspicious	severe_hypok	0.7129
RTVAE (CORE)	suspicious	comp_shock	0.61
RTVAE (CORE)	suspicious	low_cr	0.4458
RTVAE (CORE)	impossible	dbp_ge_sbp	0.2758
RTVAE (CORE)	suspicious	pp_wide	0.1914
RTVAE (CORE)	suspicious	comp_hyperk	0.1667

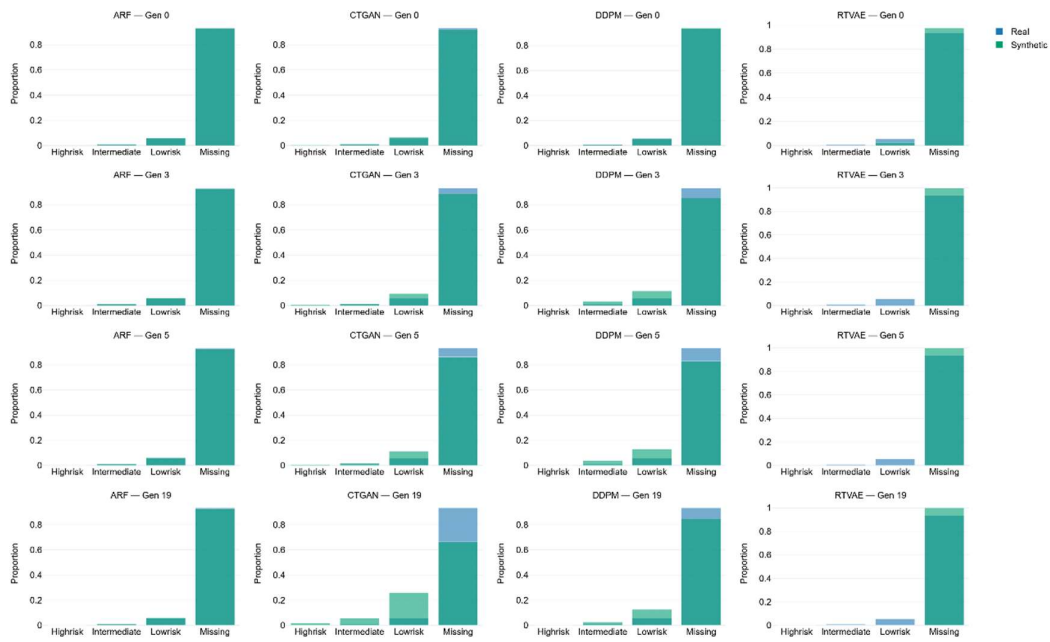
2 Supplementary data

2.1 Frequency plots for synthetic datasets

This section contains frequency plots of selected features to illustrate collapse modes and degeneration mechanics for the four generators, for generation 0, 3, 5 and 19. All are from the full dataset (high complexity) and pooled across the 15 recursive chains.

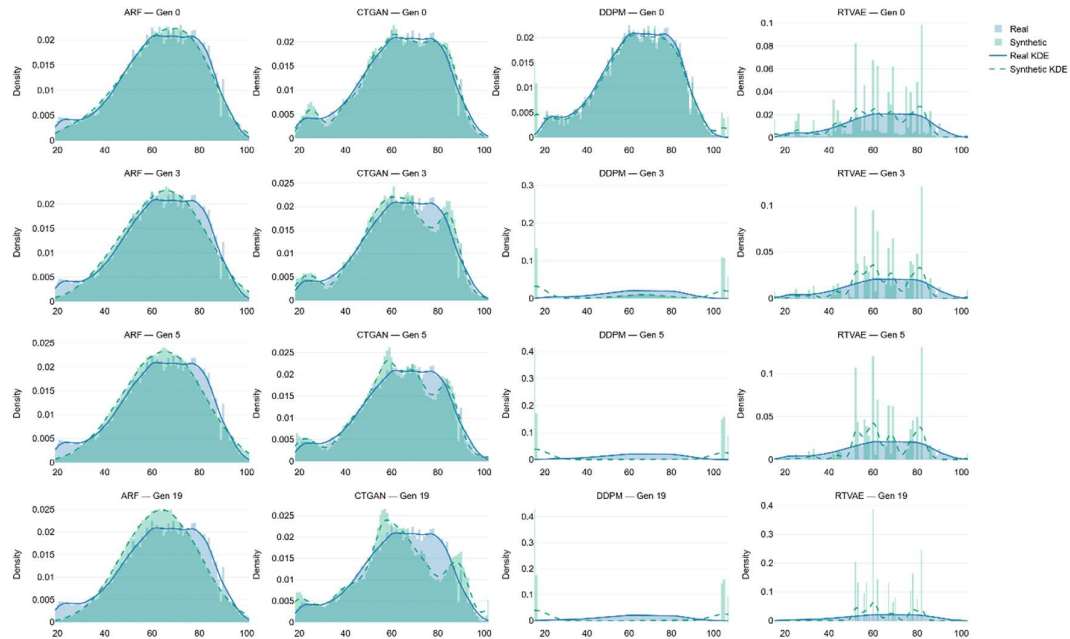


Supplementary figure 4 Readmission - 30-day readmission to ICU frequency plots over generations 0,3, 5 and 19 for real versus synthetic data (pooled over 15 chains).

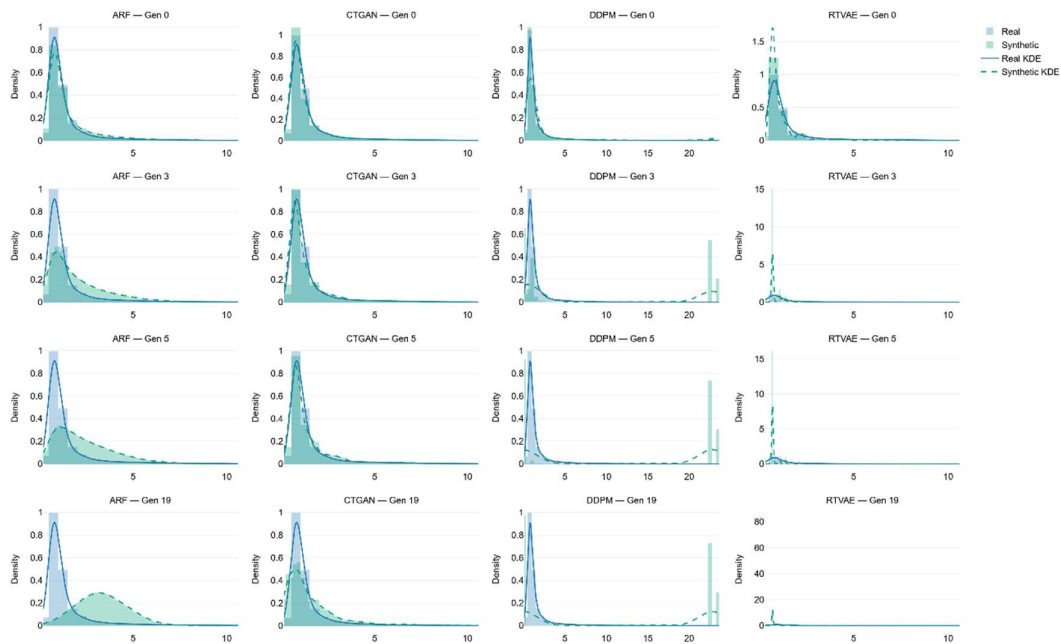


Supplementary figure 5 Cholesterol measurements, categorized.

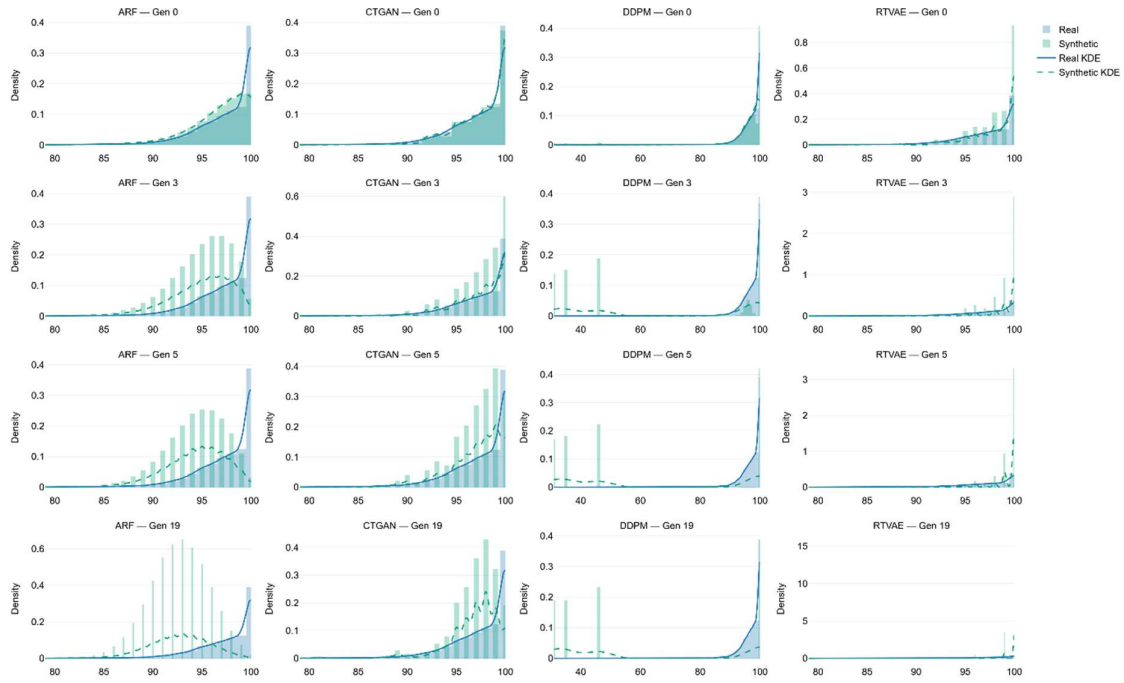
Highrisk: >300 mg/dL Intermediate: 200-300 mg/dL Lowrisk: 0-200 mg/dL frequency plots over generations 0, 3, 5 and 19 for real versus synthetic data (pooled over 15 chains).



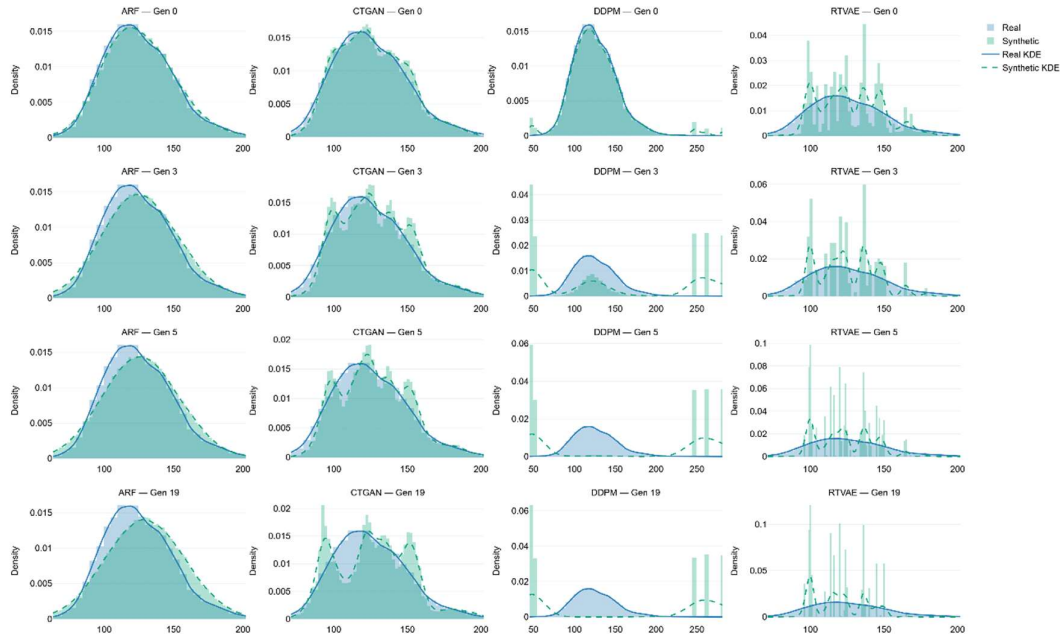
Supplementary figure 6 Age frequency plot over generations - from generation 0, 3, 5 and 19, for real versus synthetic data (pooled over 15 chains).



Supplementary figure 7 Creatinine frequency plot over generations - from generation 0, 3, 5 and 19 for real versus synthetic data (pooled over 15 chains).



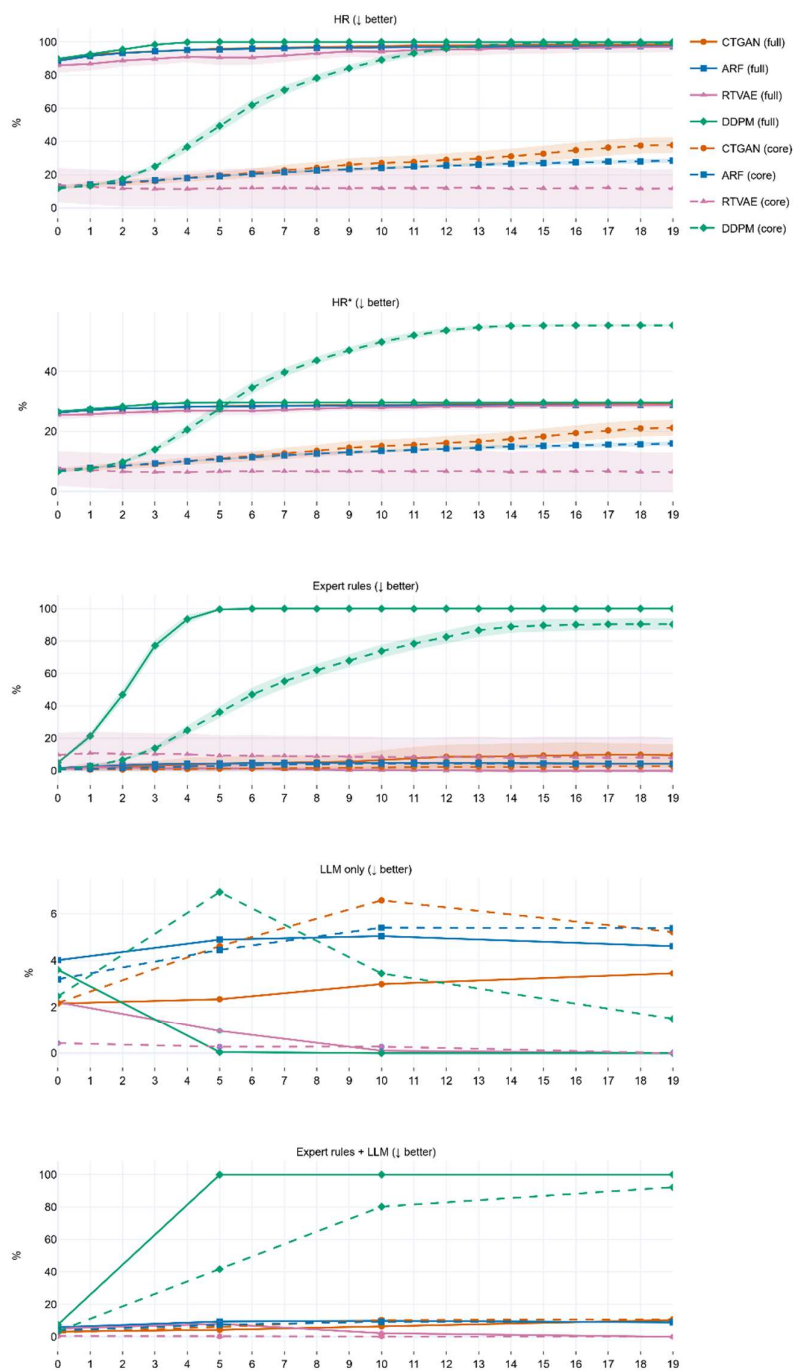
Supplementary figure 8 SPO2 frequency plot over generations - from generation 0, 3, 5 and 19 for real versus synthetic data (pooled over 15 chains).



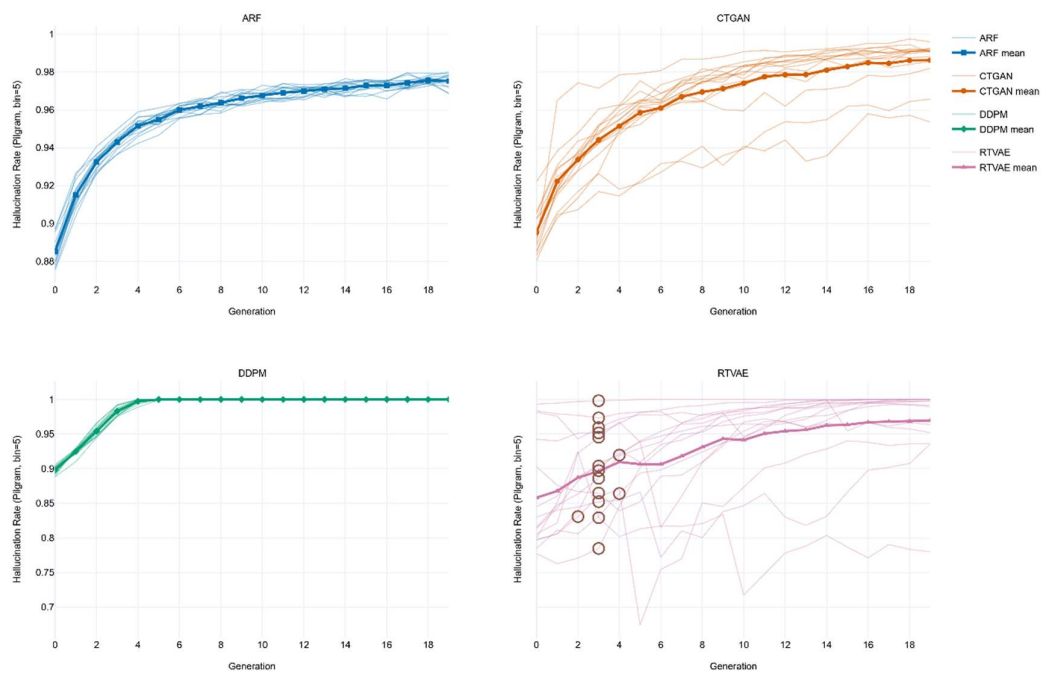
Supplementary figure 9 Systolic blood pressure (SBP) frequency plot over generations - from generation 0, 3, 5 and 19 for real versus synthetic data (pooled over 15 chains).

2.2 Evaluation metrics plots

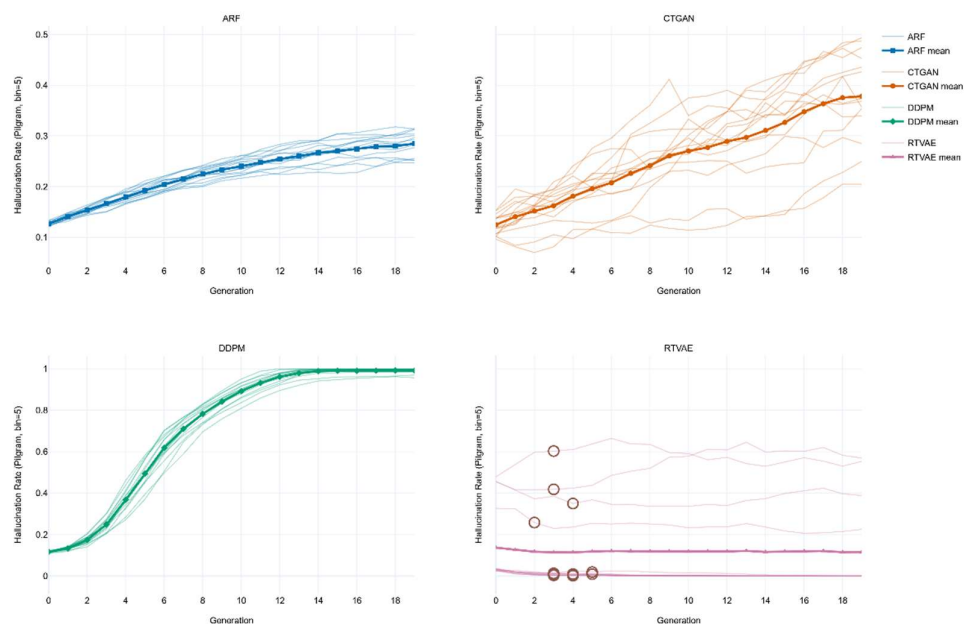
This appendix lists quality metric trajectories of results from different groups of quality evaluation metrics, for all generators and both dataset complexities. Results pooled across 15 generative chains are reported as mean values with uncertainty bands of 95 % CI. Explanation of metrics and references in supplemental methods.



Supplementary figure 10 Hallucination metrics results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands. HR and HR* are plotted for bin = 5.

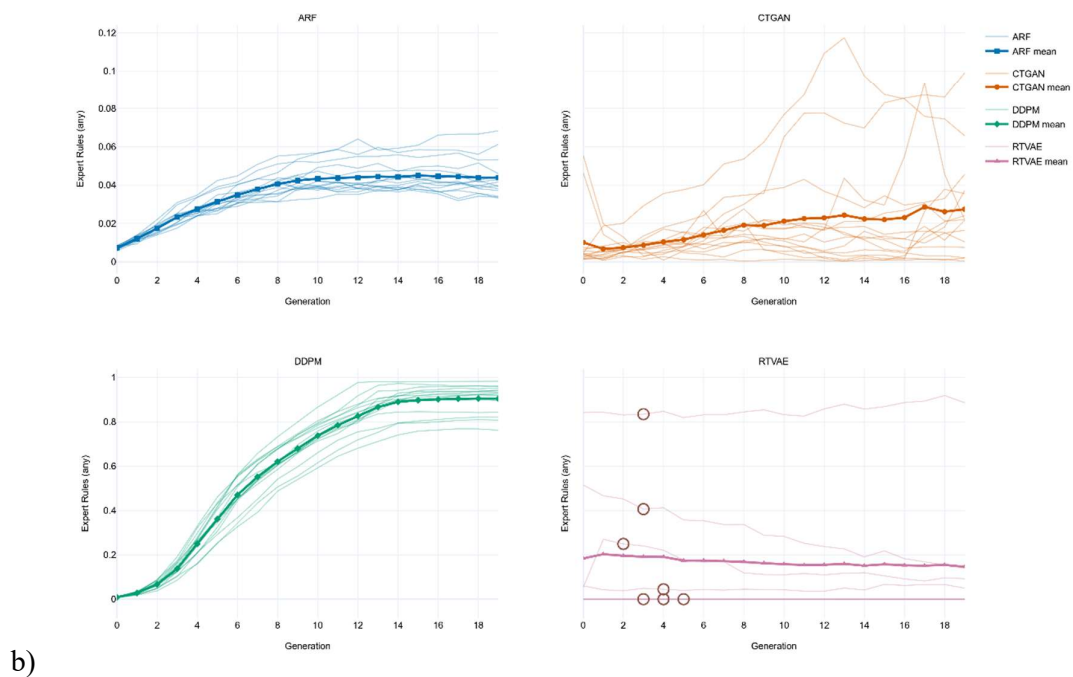
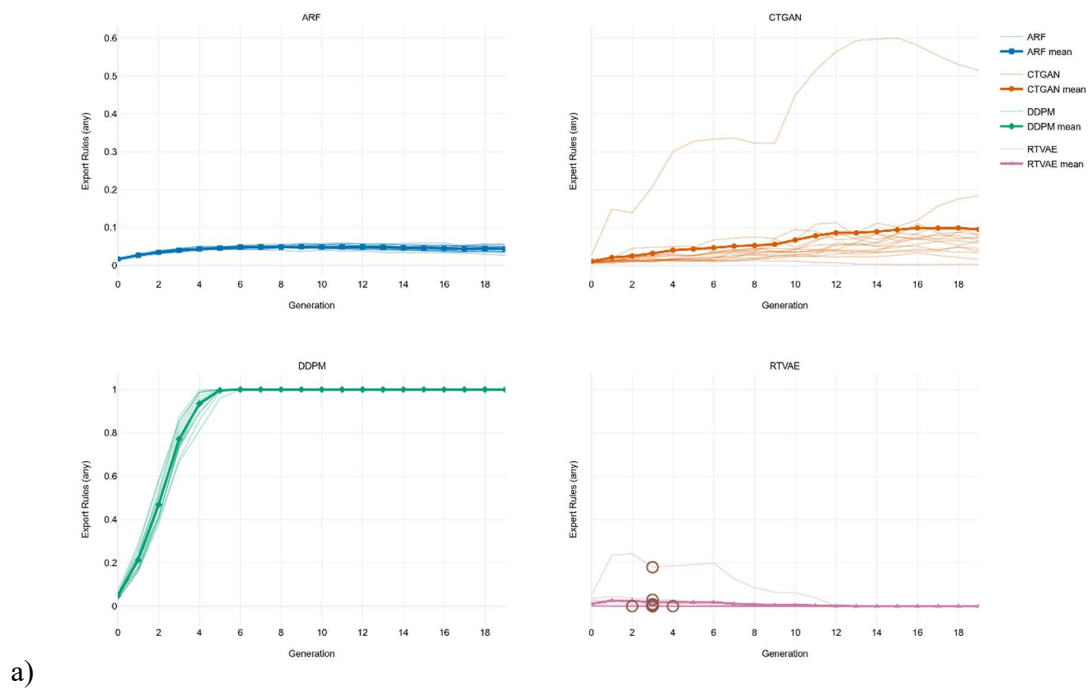


a)

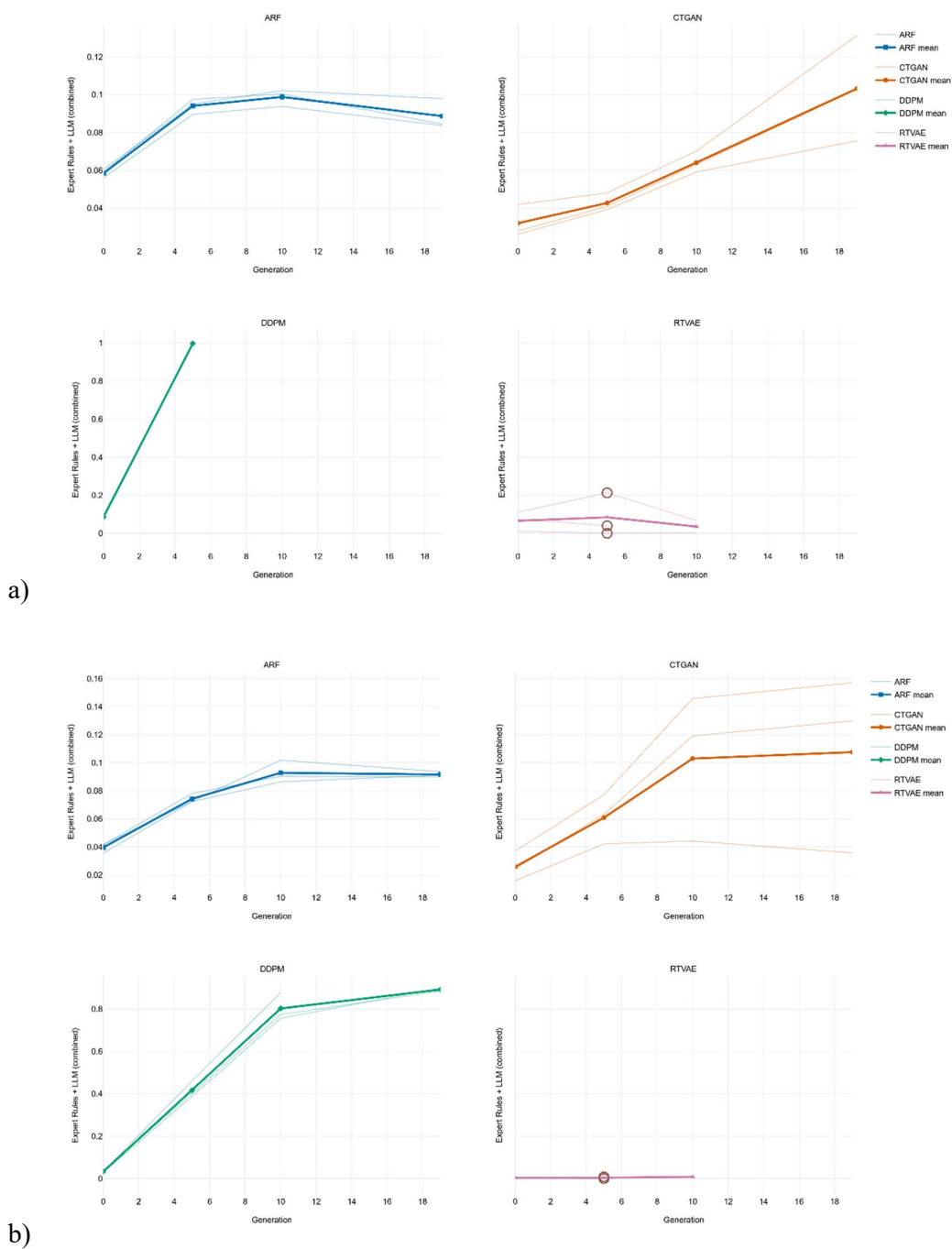


b)

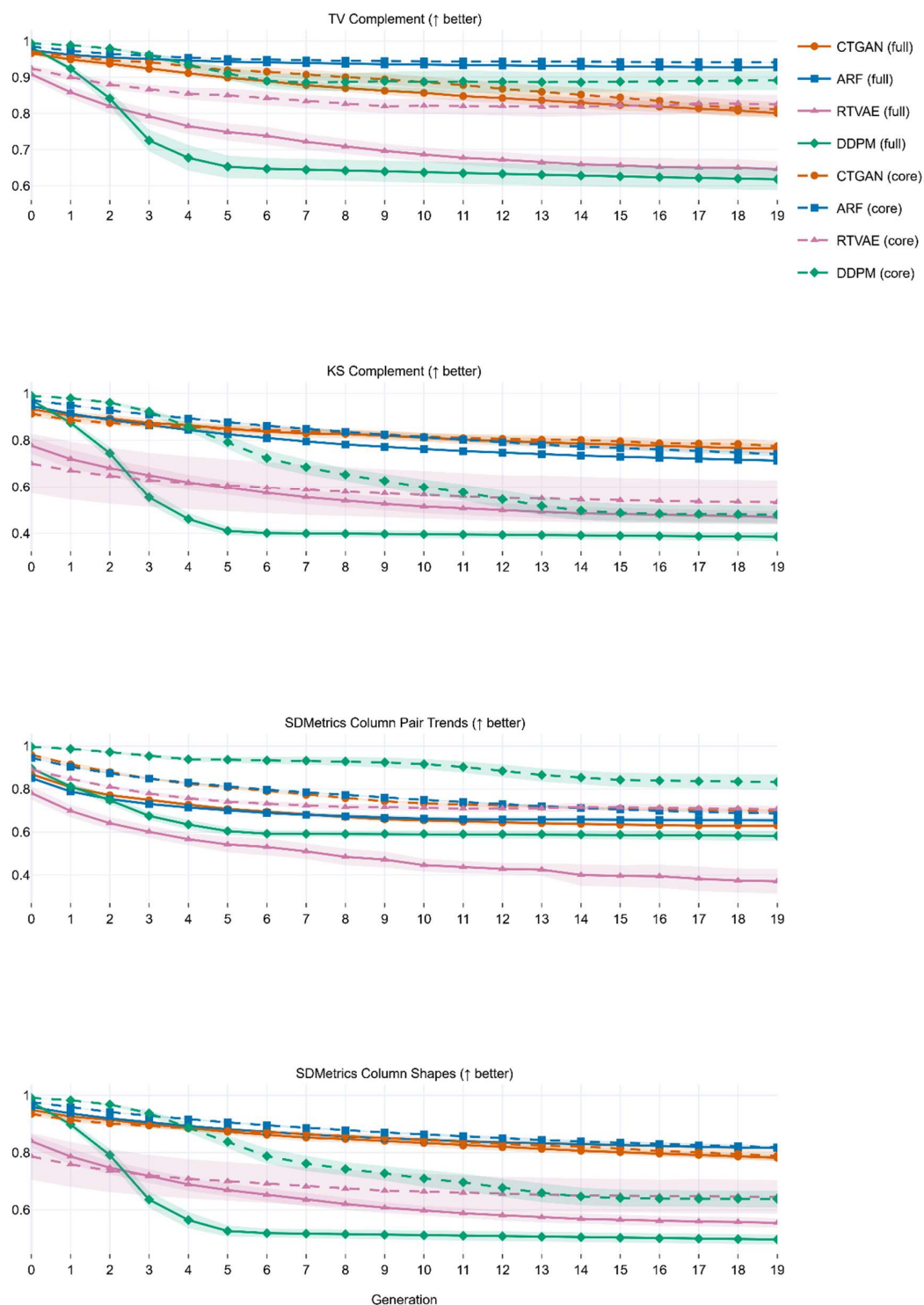
Supplementary figure 11 Per-chain plots of HR (bin=5) over 20 generations for a) full dataset and b) core dataset. Circles mark onset of readmission class collapse.



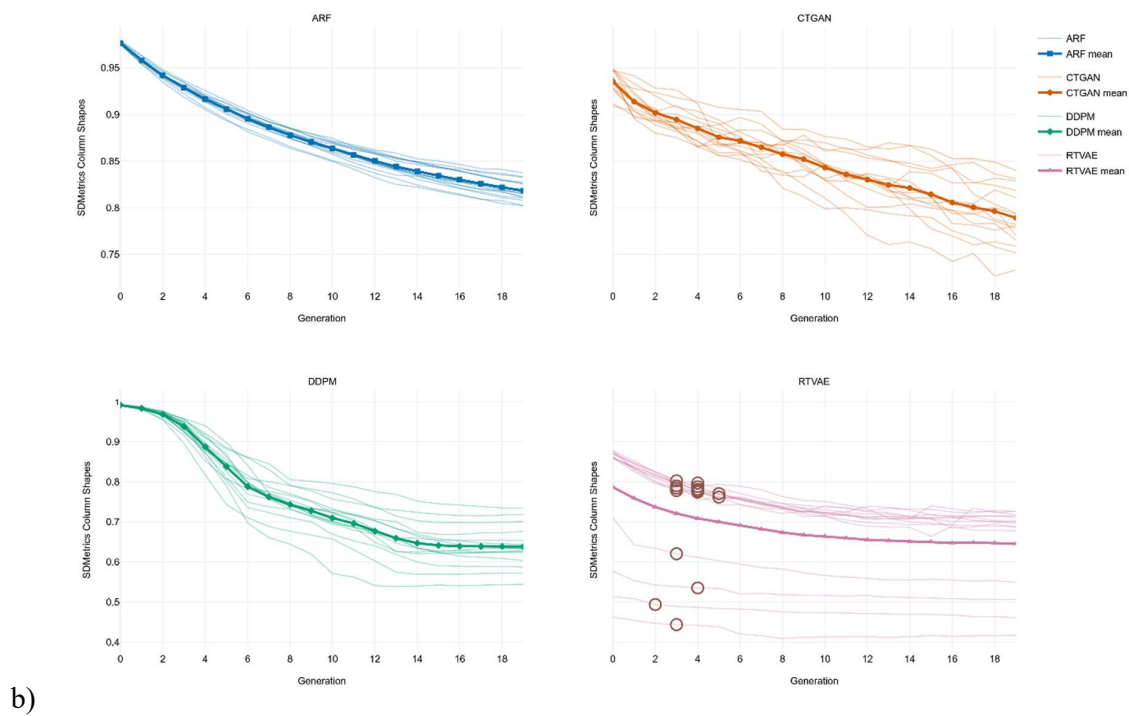
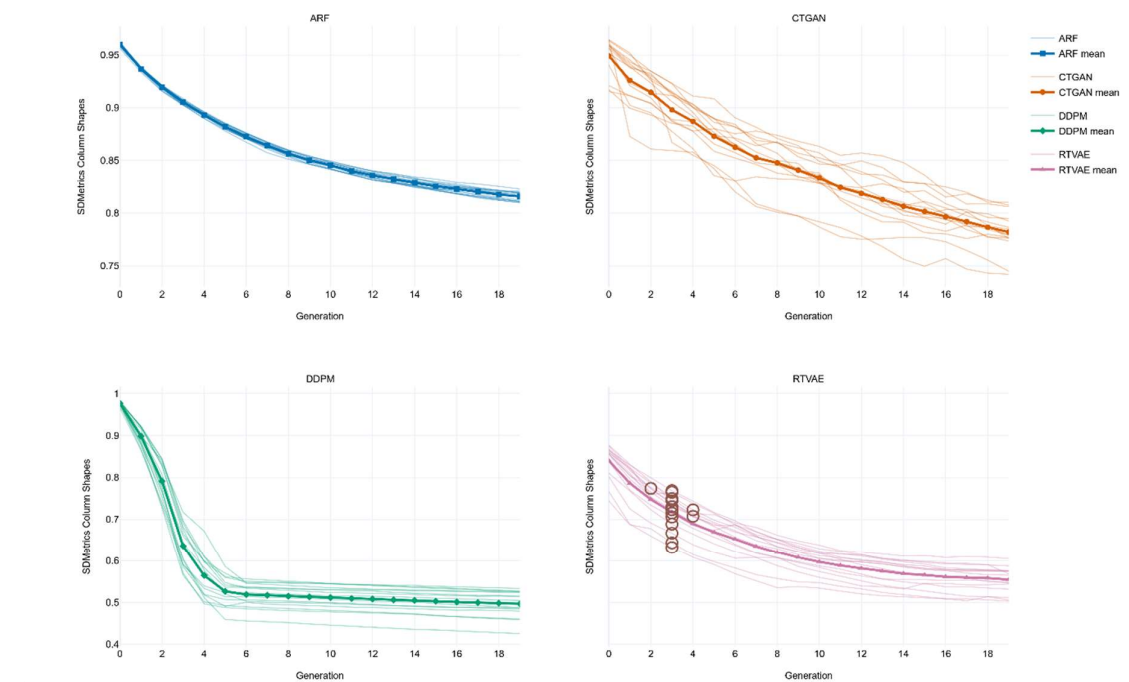
Supplementary figure 12 Per-chain plots of expert rule violations for a) full dataset and b) core dataset. Circles mark onset of readmission class collapse.



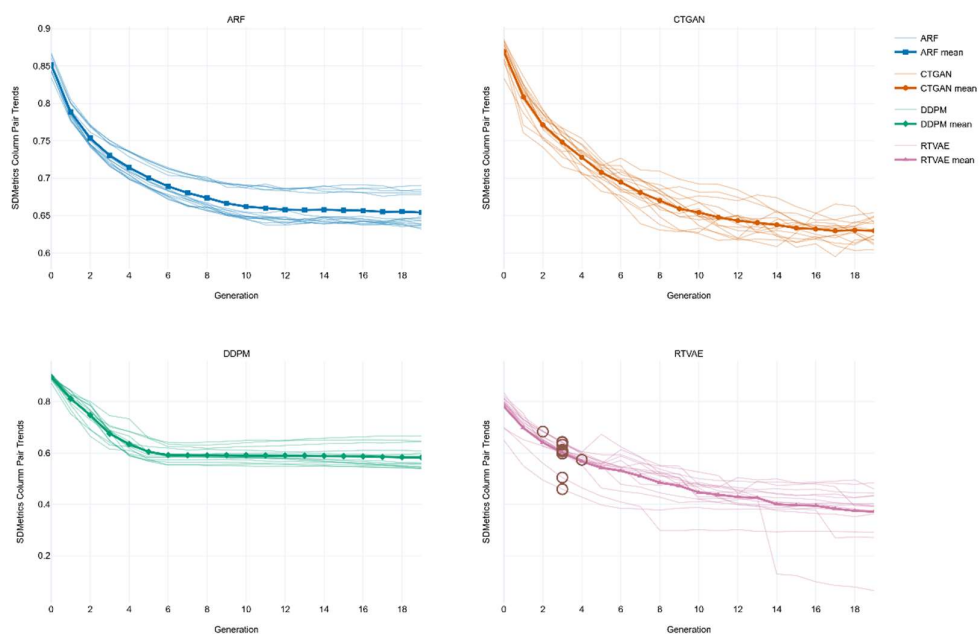
Supplementary figure 13 Per-chain plot of the combined violation rate from expert rules and LLM evaluator for 3 chains for a) full dataset and b) core dataset, over the timespan of analysed generations.



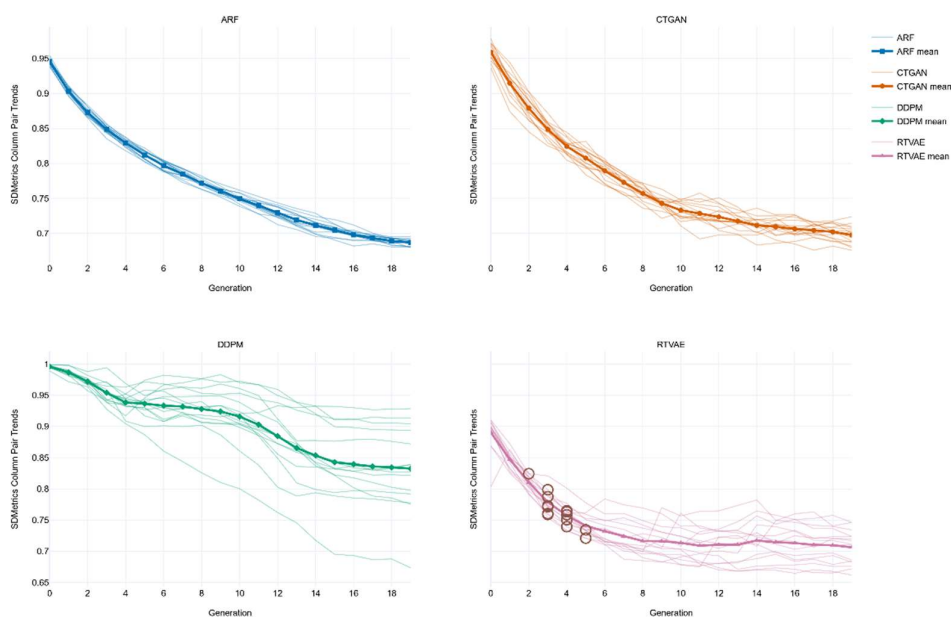
Supplementary figure 14 SD metrics results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



Supplementary figure 15 Per-chain plot of aggregate similarity (column shapes, SDV) over 20 generations for a) full dataset and b) core dataset. Circle marks onset of readmission class collapse.

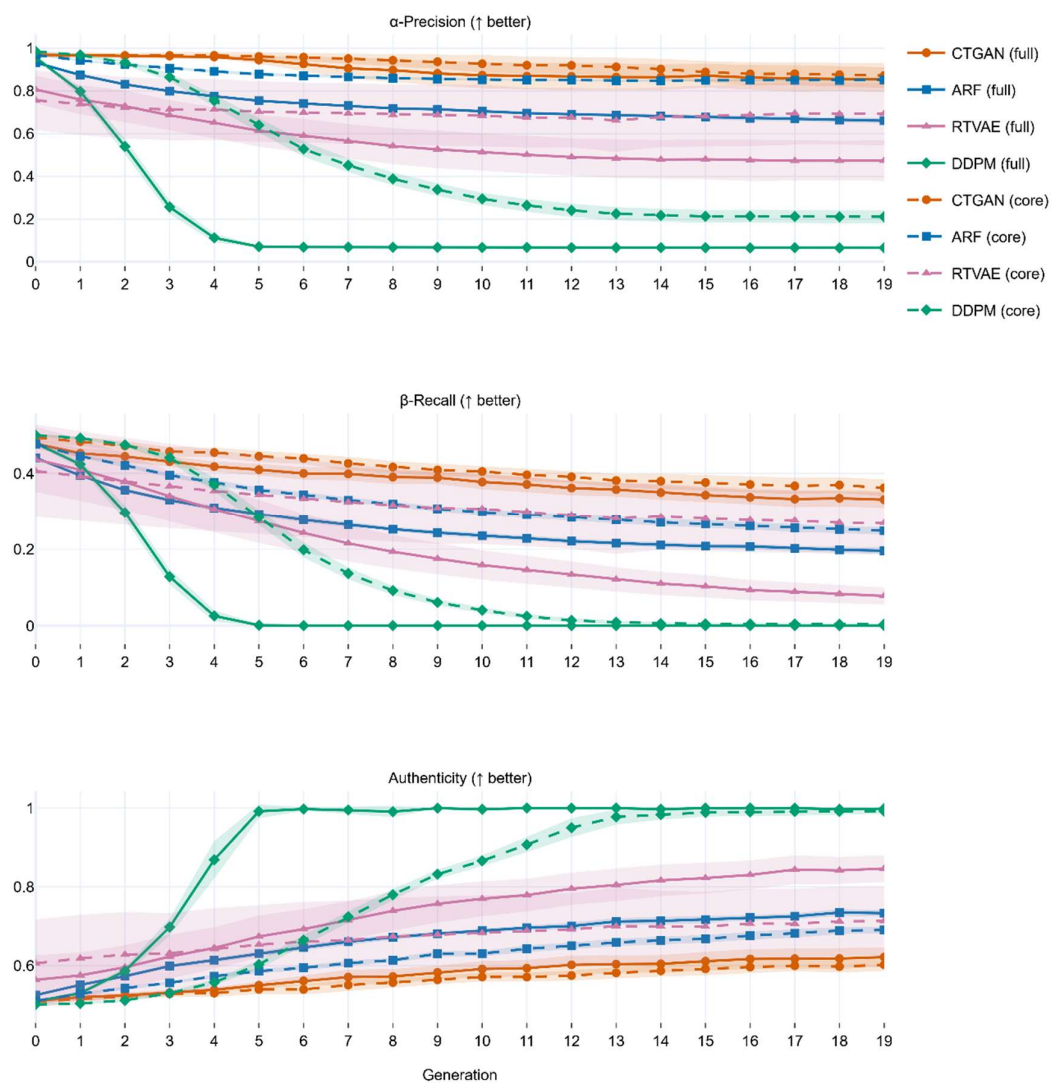


a)

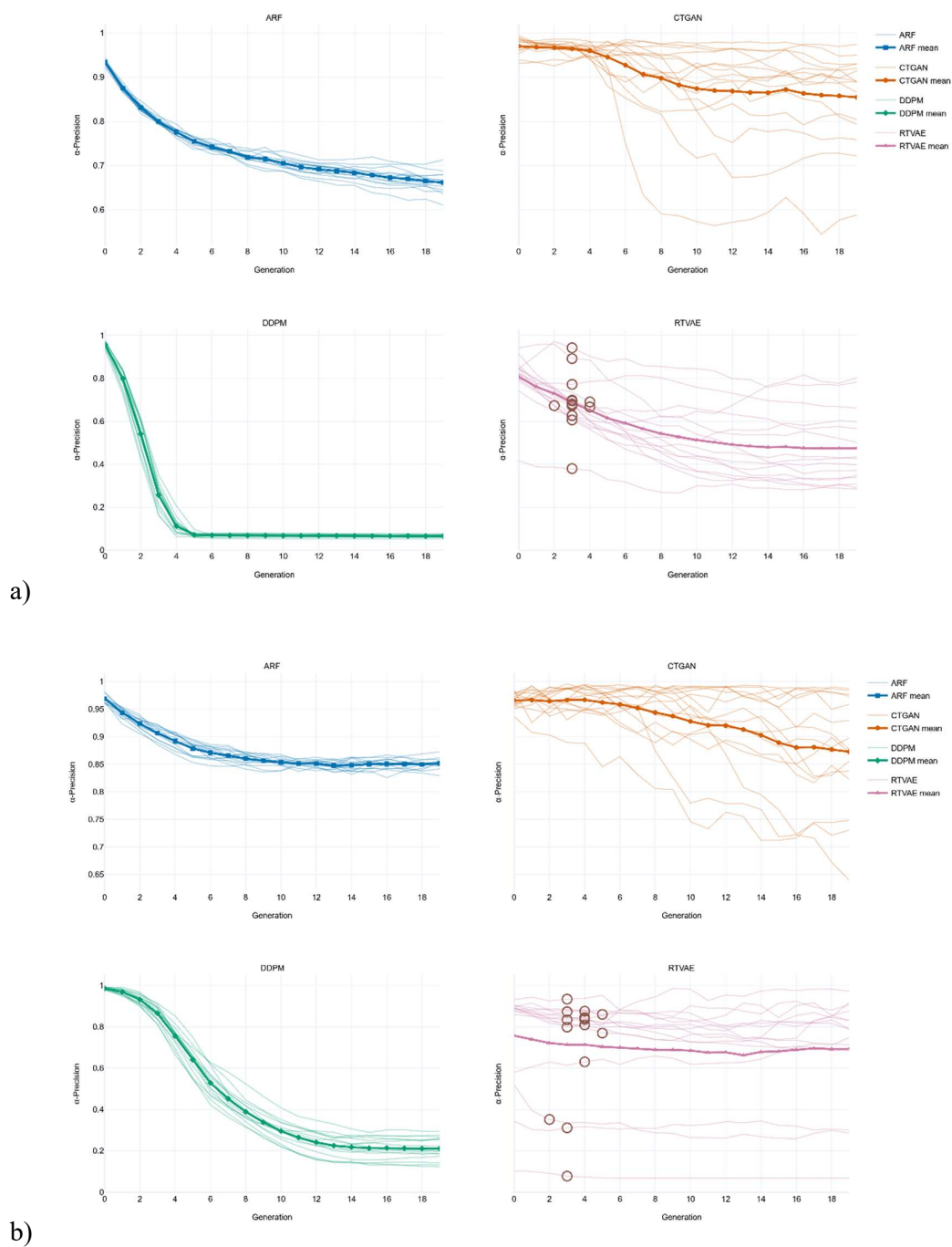


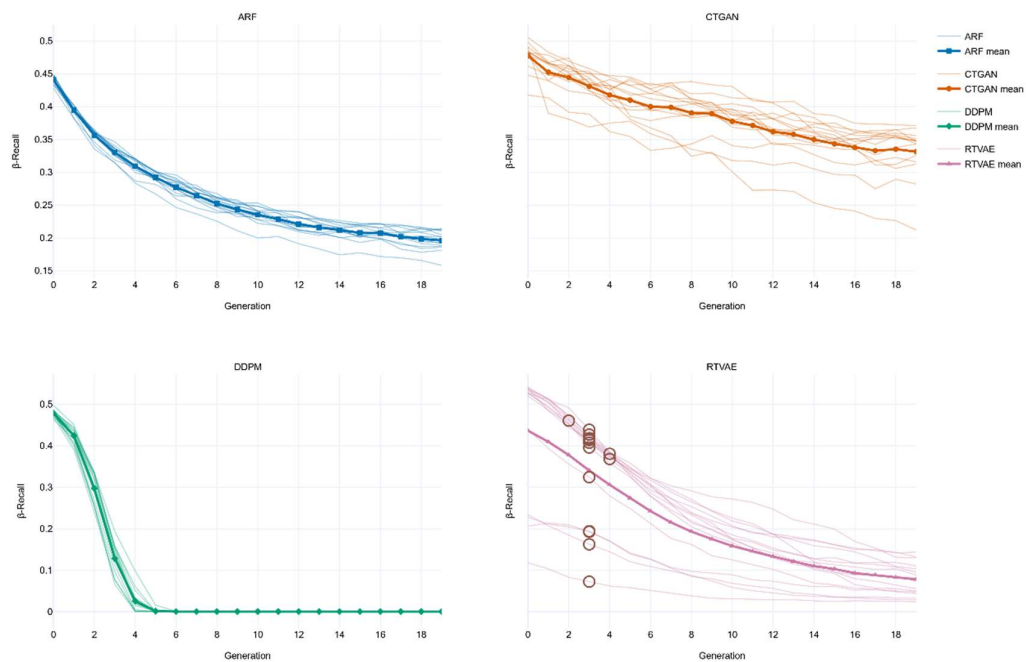
b)

Supplementary figure 16 Per-chain plot of correlation (column pair trend, SDV) for a) full dataset and b) core dataset. Circle marks onset of readmission class collapse.

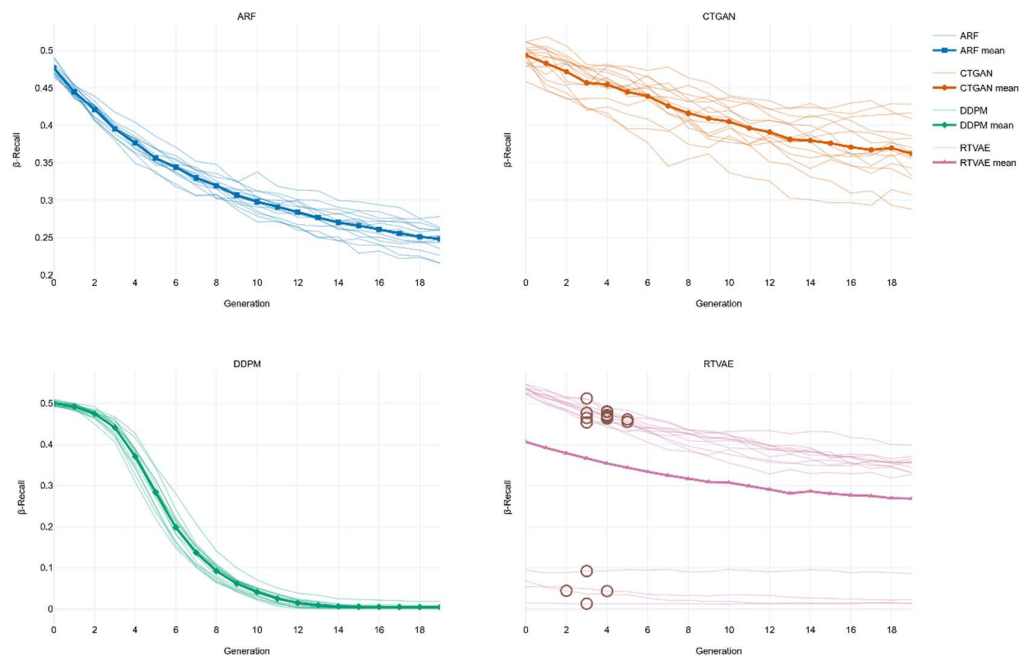


Supplementary figure 17 Fidelity, diversity and authenticity metrics of α -precision, β -recall and authenticity results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



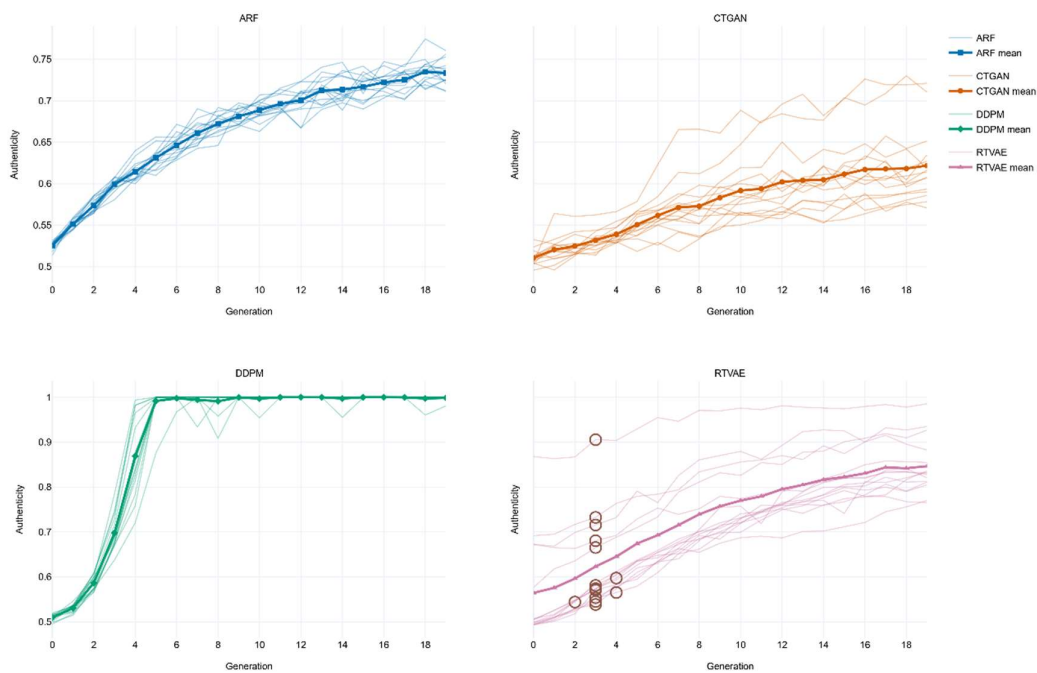


a)

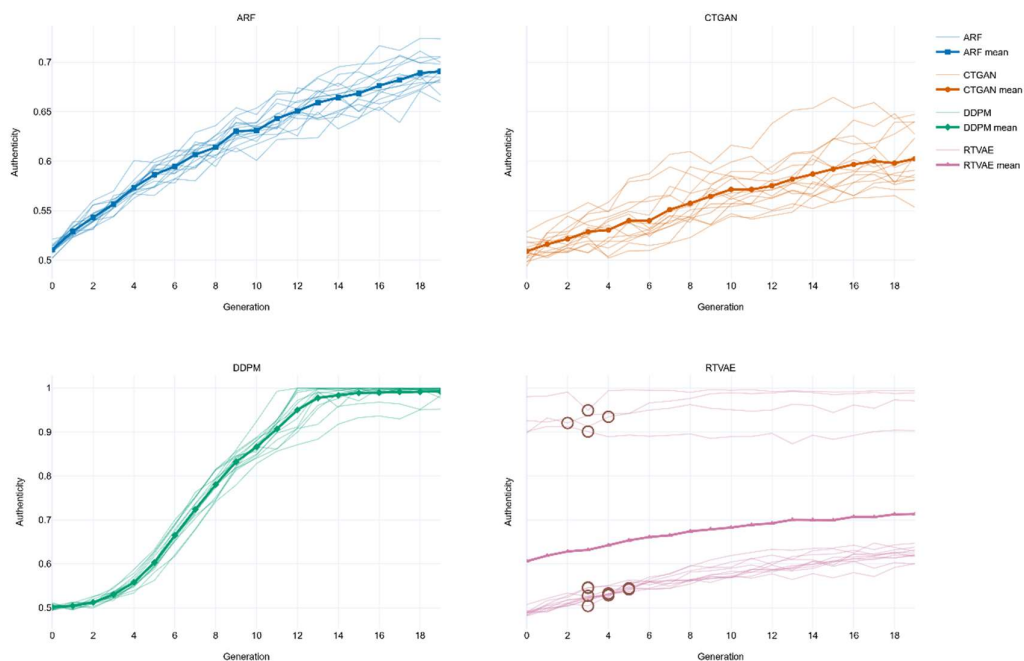


b)

Supplementary figure 19 Per-chain plot of beta-recall for a) full dataset and b) core dataset. Circle marks onset of readmission class collapse.

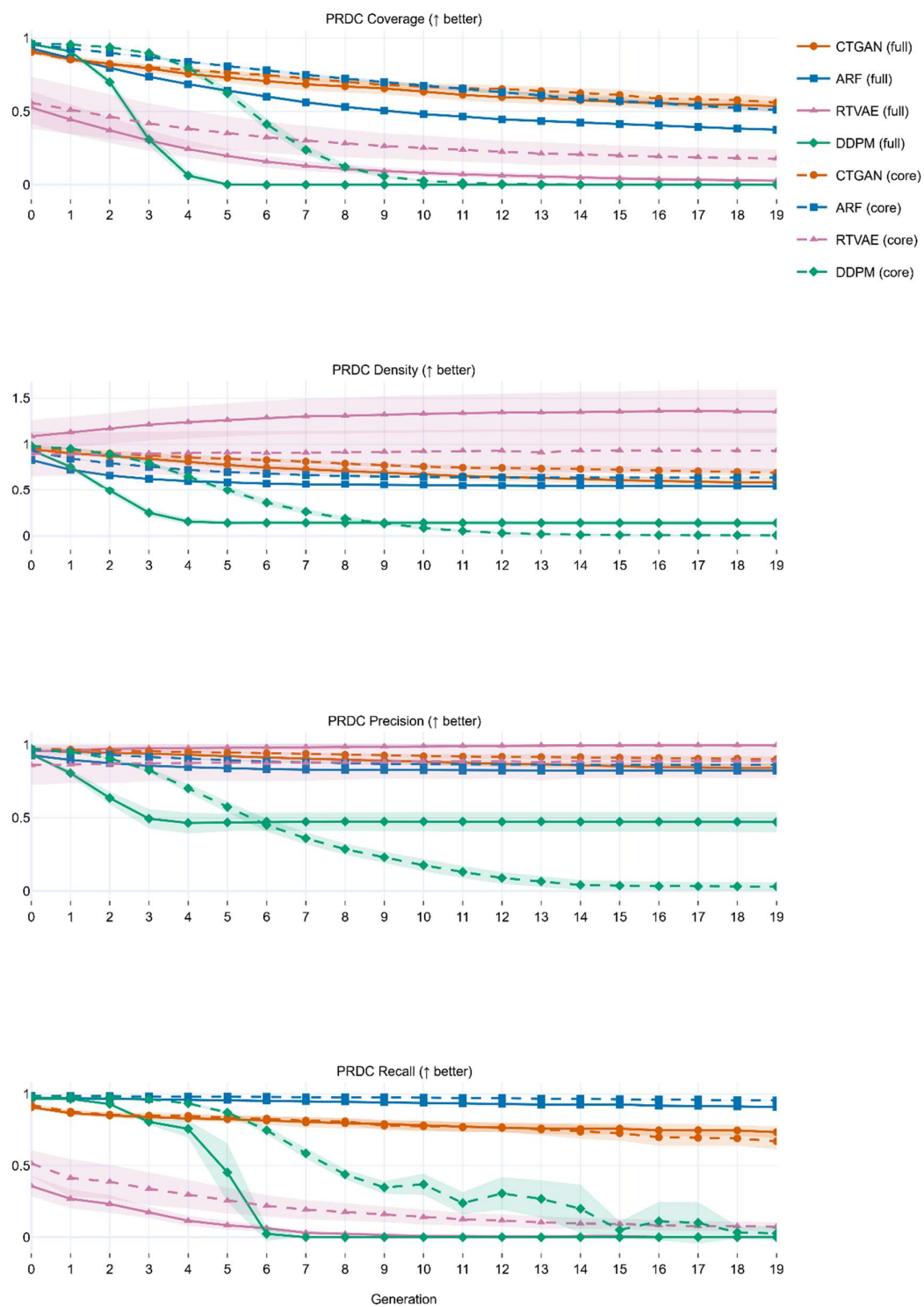


a)

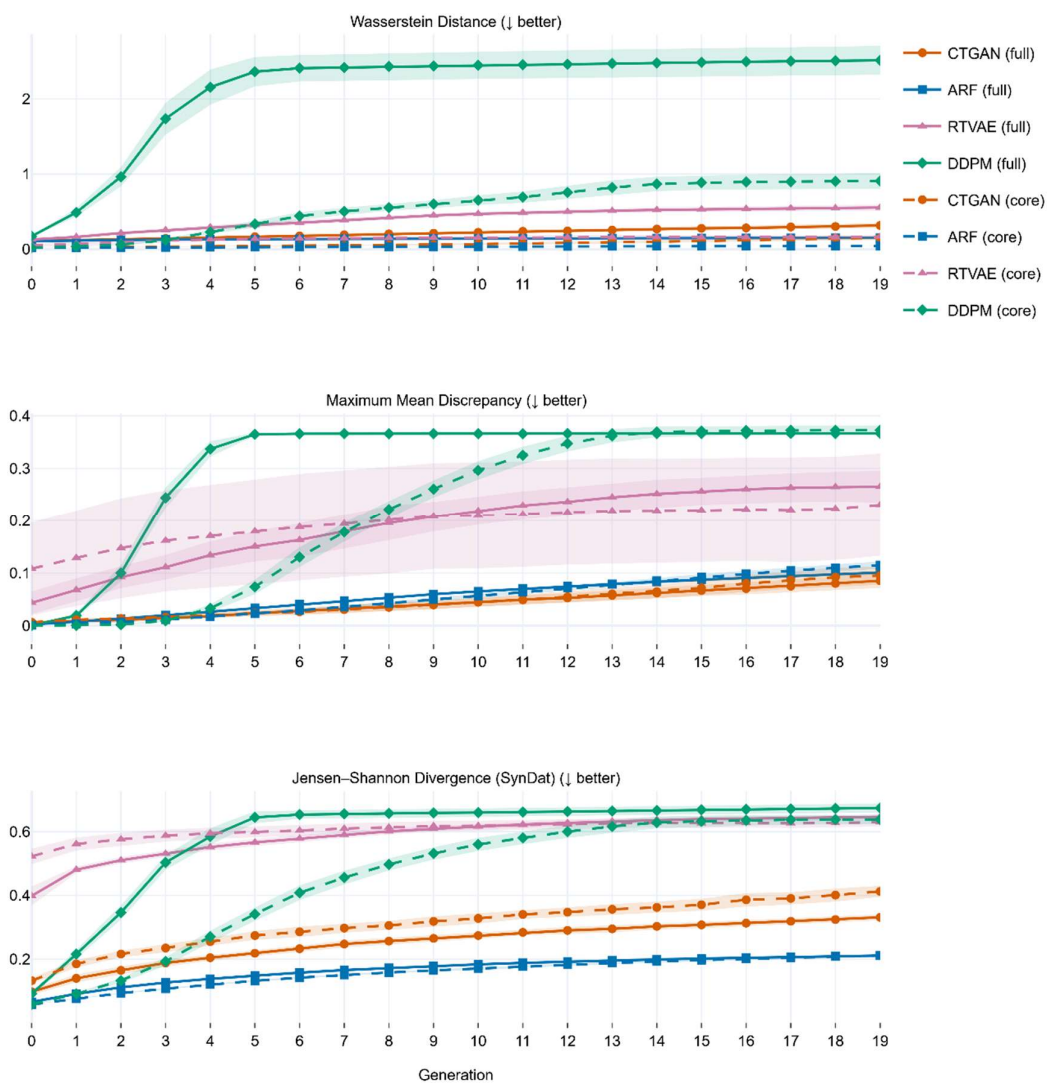


b)

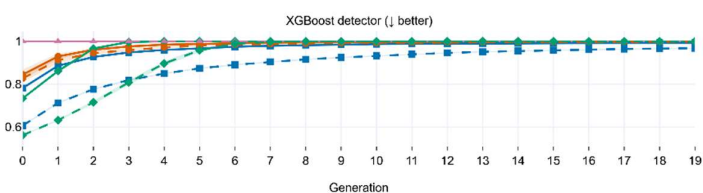
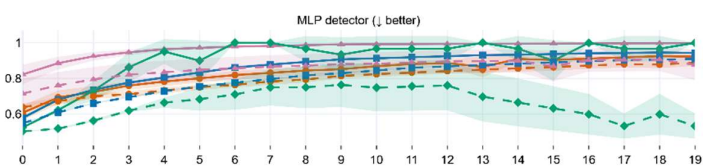
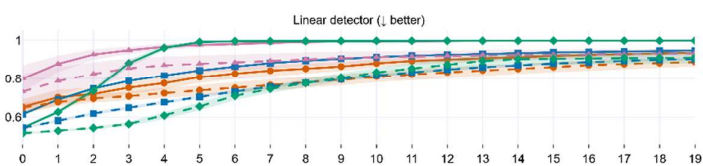
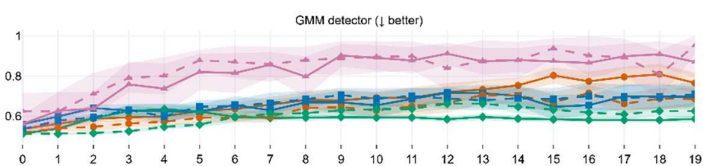
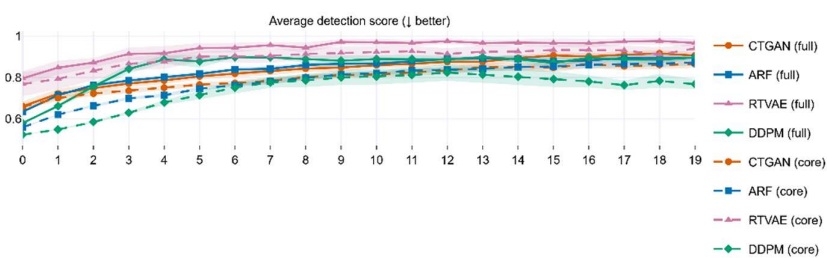
Supplementary figure 20 Per-chain plot of Authenticity for a) full dataset, b) core dataset. Circle marks onset of readmission class collapse



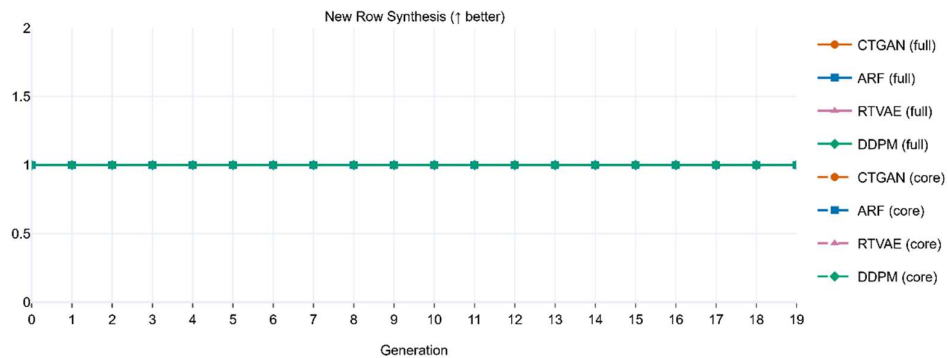
Supplementary figure 21 PRDC metrics results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



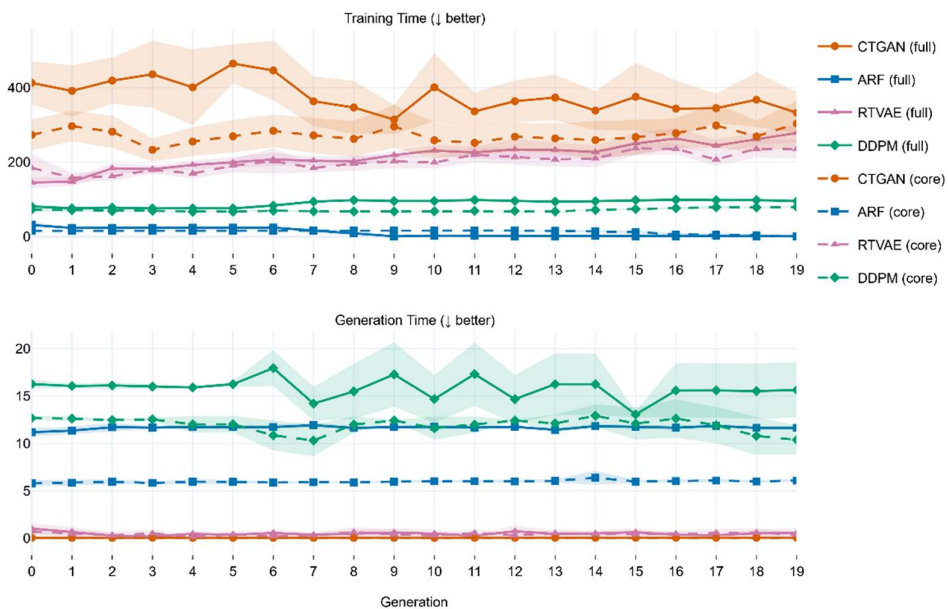
Supplementary figure 22 Aggregate distributional distance metrics results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



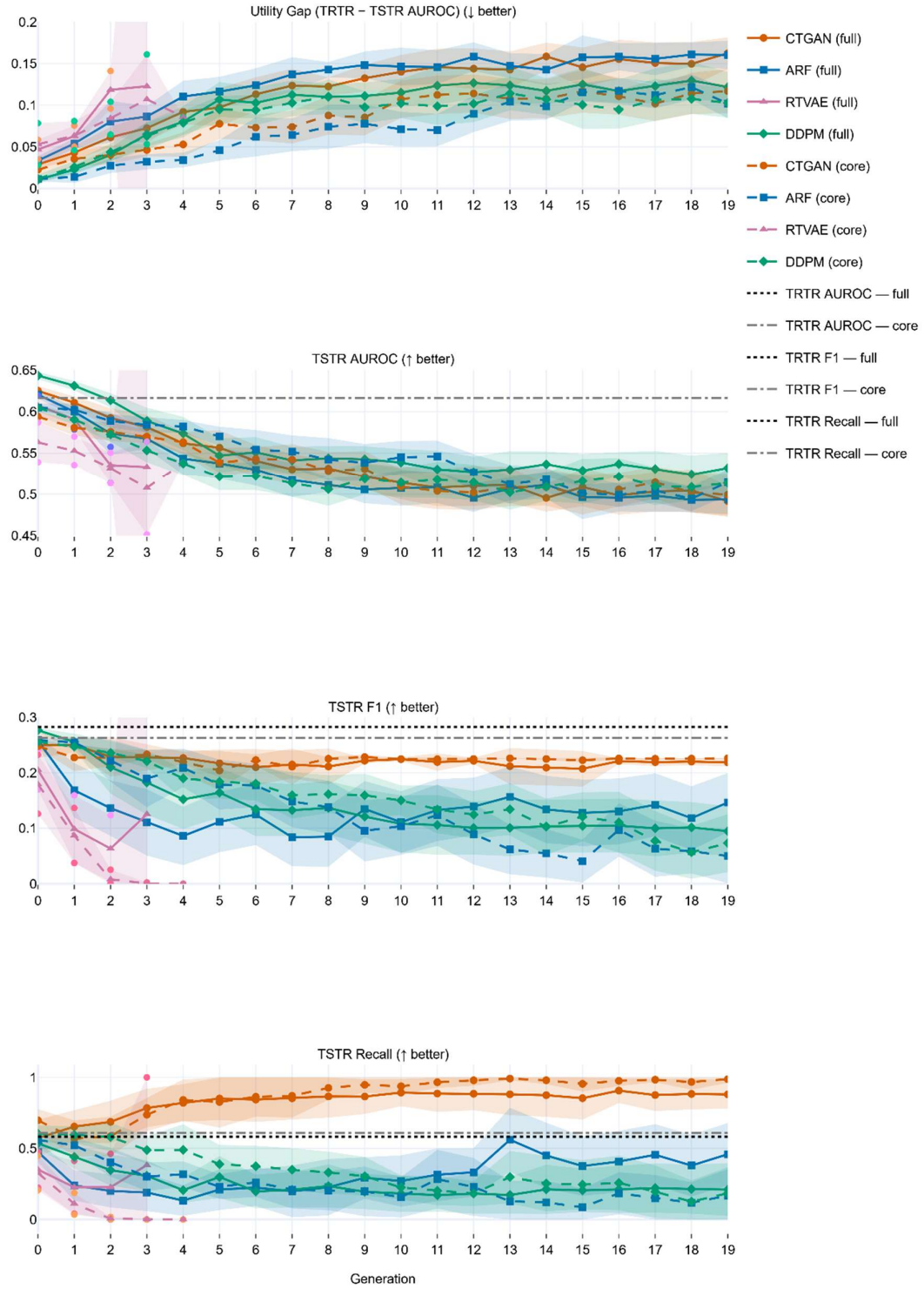
Supplementary figure 23 Binary detection metrics results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



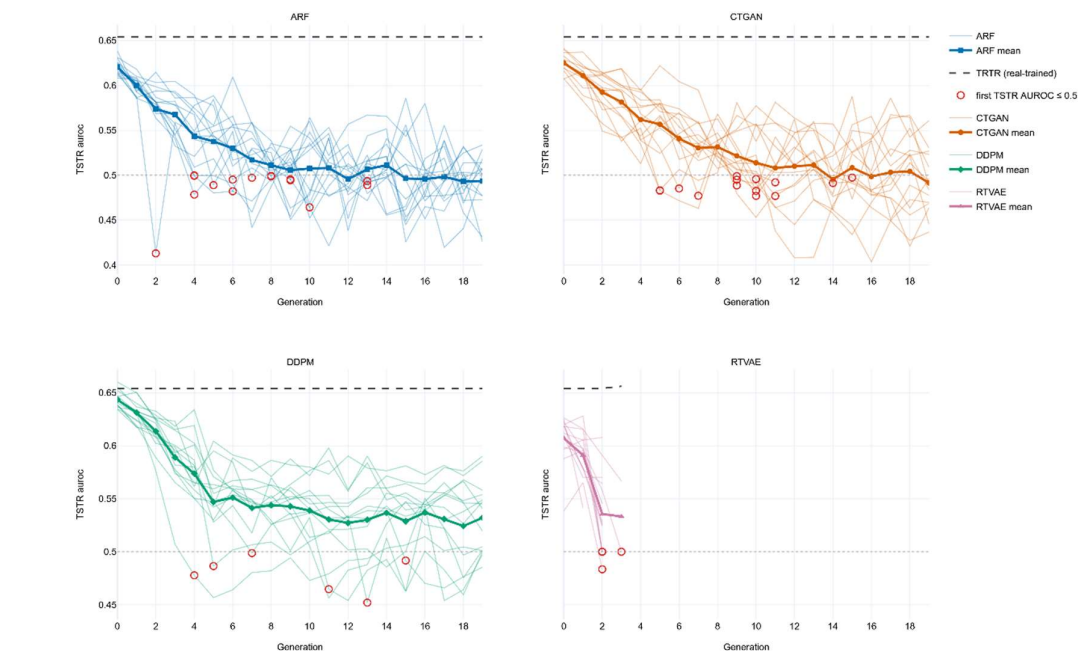
Supplementary figure 24 New row synthesis (copies) results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



Supplementary figure 25 Training time and generation time results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



Supplementary figure 26 TSTR-TRTR results pooled over generations for 15 chains. Mean values with 95% CI uncertainty bands.



a)



b)

Supplementary figure 27 TSTR AUROC results for all 15 chains plotted over 20 generations for a) the full datasets and b) the core datasets. The first time a chain collapses below 0.5 is marked with a red ring. The bold line is the mean value, and the dotted line represents the real data reference TRTR AUROC.

2.3 Results tables from quality evaluation and hallucination rates

This appendix lists tables of results from different groups of quality evaluation metrics and hallucination metrics for all generators and both dataset complexities. Results are reported as mean values with uncertainty bands of 95 % CI. Explanation of metrics and references in supplemental methods.

Table 1 Hallucination rates across recursive generations.

Reported values are the means $\pm$ half-width of 95 % CI, as a percentage of synthetic rows (10,000 per file) across 15 chains of recursive generations from the different data seed – model seed combinations for all generators and the two versions of the dataset – full variables and core variables. “impossible-fraction” means the fraction of the expert rules that were triggered by impossibility rules rather than “suspicious” rules (mean impossible / mean all expert rule violations*100). Confidence intervals are two-sided t intervals on $n-1$ degrees of freedom computed over replicate proportions, i.e. intervals on the mean of replicate proportions rather than binomial intervals on a single proportion; at extreme rates they may extend beyond [0, 100] and are reported as computed. HR is the Pilgram hallucination rate at the stated bin resolution; HR* is the entropy-adjusted rate. Expert rules (all) is the proportion of rows flagged by at least one of the expert rules; % impossible expert rules is the impossible-bucket rate as a share of that proportion, not of all rows. LLM only is the incremental contribution of the LLM evaluator: rows already flagged by the expert rules were removed before review, so the two are disjoint by construction and Expert rules + LLM is their union. LLM review was applied to generations 0, 5 and 19 only. Blank cells indicate that they are not evaluated. LLM-only and expert-plus-LLM rates are reported as the three individual per-chain values (median (min, max)). All other results are mean $\pm$ 95% CI over 15 chains. Where every replicate returns an identical value the interval collapses to zero width; this indicates metric saturation rather than precision.

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR (bin=5)	ARF	88.5% $\pm$ 0.4%	91.5% $\pm$ 0.4%	93.2% $\pm$ 0.2%	94.3% $\pm$ 0.2%	95.1% $\pm$ 0.2%	95.5% $\pm$ 0.2%	96.0% $\pm$ 0.1%	96.2% $\pm$ 0.2%	96.4% $\pm$ 0.2%	96.6% $\pm$ 0.2%	96.7% $\pm$ 0.1%	97.3% $\pm$ 0.1%	97.5% $\pm$ 0.2%
HR* (bin=5)	ARF	26.2% $\pm$ 0.1%	27.1% $\pm$ 0.1%	27.6% $\pm$ 0.1%	27.9% $\pm$ 0.1%	28.2% $\pm$ 0.1%	28.3% $\pm$ 0.1%	28.4% $\pm$ 0.0%	28.5% $\pm$ 0.0%	28.5% $\pm$ 0.1%	28.6% $\pm$ 0.1%	28.7% $\pm$ 0.0%	28.8% $\pm$ 0.0%	28.9% $\pm$ 0.1%
HR (bin=7)	ARF	96.6% $\pm$ 0.3%	98.0% $\pm$ 0.2%	98.8% $\pm$ 0.1%	99.2% $\pm$ 0.1%	99.4% $\pm$ 0.0%	99.6% $\pm$ 0.0%	99.6% $\pm$ 0.0%	99.7% $\pm$ 0.0%	99.7% $\pm$ 0.0%	99.7% $\pm$ 0.0%	99.8% $\pm$ 0.0%	99.8% $\pm$ 0.0%	99.9% $\pm$ 0.0%
HR* (bin=7)	ARF	25.5% $\pm$ 0.1%	25.9% $\pm$ 0.1%	26.1% $\pm$ 0.0%	26.2% $\pm$ 0.0%	26.3% $\pm$ 0.0%	26.3% $\pm$ 0.0%	26.3% $\pm$ 0.0%	26.3% $\pm$ 0.0%	26.4% $\pm$ 0.0%	26.4% $\pm$ 0.0%	26.4% $\pm$ 0.0%	26.4% $\pm$ 0.0%	26.4% $\pm$ 0.0%
HR (bin=10)	ARF	99.4% $\pm$ 0.1%	99.7% $\pm$ 0.0%	99.9% $\pm$ 0.0%	99.9% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
HR* (bin=10)	ARF	23.6% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%	23.7% $\pm$ 0.0%
HR (bin=15)	ARF	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
HR* (bin=15)	ARF	21.2% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%	21.3% $\pm$ 0.0%
HR (bin=20)	ARF	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%
HR* (bin=20)	ARF	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%	19.8% $\pm$ 0.0%
HR (bin=30)	ARF	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%	100.0% $\pm$ 0.0%

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR* (bin=30)	ARF	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%
Expert rules (all) %	ARF	1.7% ± 0.1%	2.7% ± 0.1%	3.4% ± 0.2%	4.0% ± 0.2%	4.3% ± 0.2%	4.6% ± 0.2%	4.8% ± 0.2%	4.8% ± 0.2%	4.8% ± 0.2%	4.9% ± 0.3%	4.9% ± 0.3%	4.6% ± 0.4%	4.3% ± 0.5%
% impossible expert rules	ARF	72.00%	76.90%	78.70%	80.80%	82.60%	84.40%	85.10%	86.90%	88.10%	88.50%	88.90%	90.20%	90.70%
LLM only %	ARF	4.1% (3.8%–4.1%)					4.8% (4.6%–5.3%)					5.0% (4.7%–5.5%)		4.8% (4.3%–4.8%)
Expert rules + LLM %	ARF	5.9% (5.6%–6.0%)					9.5% (9.0%–9.8%)					10.0% (9.4%–10.2%)		8.4% (8.4%–9.8%)
HR (bin=5)	ARF CORE	12.7% ± 0.2%	14.1% ± 0.3%	15.4% ± 0.3%	16.7% ± 0.5%	17.9% ± 0.5%	19.2% ± 0.6%	20.4% ± 0.6%	21.5% ± 0.6%	22.5% ± 0.7%	23.3% ± 0.7%	24.1% ± 0.8%	27.0% ± 1.2%	28.5% ± 1.3%
HR* (bin=5)	ARF CORE	7.1% ± 0.1%	7.9% ± 0.2%	8.6% ± 0.2%	9.3% ± 0.3%	10.0% ± 0.3%	10.7% ± 0.3%	11.4% ± 0.4%	12.0% ± 0.4%	12.6% ± 0.4%	13.0% ± 0.4%	13.4% ± 0.5%	15.1% ± 0.7%	15.9% ± 0.7%
HR (bin=7)	ARF CORE	27.4% ± 0.2%	30.4% ± 0.4%	33.2% ± 0.4%	35.8% ± 0.5%	38.6% ± 0.8%	40.9% ± 0.8%	43.1% ± 0.8%	45.4% ± 1.0%	47.4% ± 1.0%	49.2% ± 0.9%	50.9% ± 1.0%	57.1% ± 1.3%	60.3% ± 1.3%
HR* (bin=7)	ARF CORE	13.4% ± 0.1%	14.9% ± 0.2%	16.3% ± 0.2%	17.5% ± 0.2%	18.9% ± 0.4%	20.0% ± 0.4%	21.1% ± 0.4%	22.3% ± 0.5%	23.2% ± 0.5%	24.1% ± 0.4%	24.9% ± 0.5%	28.0% ± 0.7%	29.6% ± 0.6%
HR (bin=10)	ARF CORE	51.3% ± 0.3%	56.2% ± 0.4%	60.2% ± 0.5%	63.8% ± 0.6%	67.1% ± 0.7%	69.8% ± 0.8%	72.2% ± 0.8%	74.6% ± 0.8%	76.8% ± 0.8%	78.6% ± 0.7%	80.2% ± 0.8%	86.6% ± 0.8%	89.3% ± 0.6%
HR* (bin=10)	ARF CORE	22.3% ± 0.1%	24.4% ± 0.2%	26.1% ± 0.2%	27.7% ± 0.3%	29.1% ± 0.3%	30.3% ± 0.3%	31.3% ± 0.3%	32.4% ± 0.3%	33.3% ± 0.4%	34.1% ± 0.3%	34.8% ± 0.3%	37.5% ± 0.3%	38.7% ± 0.2%
HR (bin=15)	ARF CORE	81.8% ± 0.3%	85.0% ± 0.3%	87.6% ± 0.3%	89.4% ± 0.4%	91.0% ± 0.3%	92.4% ± 0.3%	93.4% ± 0.3%	94.4% ± 0.2%	95.2% ± 0.3%	95.8% ± 0.3%	96.4% ± 0.2%	98.3% ± 0.1%	98.9% ± 0.1%
HR* (bin=15)	ARF CORE	31.4% ± 0.1%	32.6% ± 0.1%	33.6% ± 0.1%	34.3% ± 0.1%	34.9% ± 0.1%	35.4% ± 0.1%	35.8% ± 0.1%	36.2% ± 0.1%	36.5% ± 0.1%	36.8% ± 0.1%	37.0% ± 0.1%	37.7% ± 0.1%	37.9% ± 0.0%
HR (bin=20)	ARF CORE	94.9% ± 0.1%	96.1% ± 0.1%	97.0% ± 0.1%	97.5% ± 0.1%	98.0% ± 0.2%	98.3% ± 0.1%	98.7% ± 0.1%	98.8% ± 0.1%	99.1% ± 0.1%	99.3% ± 0.1%	99.4% ± 0.1%	99.8% ± 0.0%	99.9% ± 0.0%
HR* (bin=20)	ARF CORE	33.6% ± 0.0%	34.1% ± 0.0%	34.4% ± 0.0%	34.6% ± 0.0%	34.8% ± 0.1%	34.9% ± 0.0%	35.0% ± 0.0%	35.0% ± 0.0%	35.1% ± 0.0%	35.2% ± 0.0%	35.2% ± 0.0%	35.4% ± 0.0%	35.4% ± 0.0%
HR (bin=30)	ARF CORE	99.6% ± 0.0%	99.7% ± 0.0%	99.8% ± 0.0%	99.8% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	ARF CORE	31.9% ± 0.0%	31.9% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%
Expert rules (all) %	ARF CORE	0.7% ± 0.0%	1.2% ± 0.1%	1.7% ± 0.1%	2.3% ± 0.2%	2.7% ± 0.2%	3.1% ± 0.3%	3.5% ± 0.3%	3.8% ± 0.3%	4.1% ± 0.4%	4.2% ± 0.4%	4.3% ± 0.4%	4.5% ± 0.4%	4.4% ± 0.6%
% impossible expert rules	ARF CORE	71.00%	77.30%	80.90%	83.00%	84.70%	86.10%	87.30%	89.50%	90.00%	90.60%	91.80%	93.50%	95.20%
LLM only %	ARF CORE	3.4% (2.7%–3.5%)					4.5% (4.1%–4.8%)					5.1% (4.9%–6.2%)		5.3% (5.2%–5.6%)
Expert rules + LLM %	ARF CORE	4.2% (3.6%–4.2%)					7.2% (7.2%–7.8%)					9.0% (8.6%–10.2%)		9.1% (9.0%–9.4%)

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR (bin=5)	CTGAN	89.5% ± 0.6%	92.2% ± 0.9%	93.4% ± 0.9%	94.4% ± 0.8%	95.1% ± 0.9%	95.9% ± 0.9%	96.1% ± 0.8%	96.7% ± 0.8%	96.9% ± 0.9%	97.1% ± 0.9%	97.4% ± 0.9%	98.3% ± 0.7%	98.6% ± 0.6%
HR* (bin=5)	CTGAN	26.5% ± 0.2%	27.3% ± 0.3%	27.7% ± 0.3%	28.0% ± 0.2%	28.2% ± 0.3%	28.4% ± 0.3%	28.5% ± 0.2%	28.6% ± 0.2%	28.7% ± 0.3%	28.8% ± 0.3%	28.9% ± 0.3%	29.1% ± 0.2%	29.2% ± 0.2%
HR (bin=7)	CTGAN	97.5% ± 0.2%	98.6% ± 0.2%	99.0% ± 0.2%	99.2% ± 0.1%	99.4% ± 0.2%	99.5% ± 0.1%	99.6% ± 0.1%	99.7% ± 0.1%	99.7% ± 0.1%	99.7% ± 0.1%	99.8% ± 0.1%	99.9% ± 0.1%	99.9% ± 0.1%
HR* (bin=7)	CTGAN	25.8% ± 0.1%	26.1% ± 0.1%	26.2% ± 0.0%	26.2% ± 0.0%	26.3% ± 0.0%	26.3% ± 0.0%	26.3% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%
HR (bin=10)	CTGAN	99.6% ± 0.0%	99.8% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=10)	CTGAN	23.6% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%
HR (bin=15)	CTGAN	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=15)	CTGAN	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%
HR (bin=20)	CTGAN	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=20)	CTGAN	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%
HR (bin=30)	CTGAN	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	CTGAN	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%
Expert rules (all) %	CTGAN	1.0% ± 0.3%	2.1% ± 2.0%	2.5% ± 1.8%	3.2% ± 2.8%	4.0% ± 4.0%	4.3% ± 4.0%	4.7% ± 4.5%	5.1% ± 4.5%	5.3% ± 4.2%	5.6% ± 4.2%	6.7% ± 6.0%	9.4% ± 7.9%	9.5% ± 6.8%
% impossible expert rules	CTGAN	39.70%	23.70%	31.70%	29.10%	25.90%	31.90%	34.20%	36.80%	40.20%	42.30%	39.00%	46.00%	46.80%
LLM only %	CTGAN	2.1% (2.0%–2.4%)					2.2% (2.2%–2.6%)					2.9% (2.9%–3.1%)		3.3% (3.1%–4.0%)
Expert rules + LLM %	CTGAN	2.8% (2.6%–4.2%)					4.1% (3.9%–4.8%)					6.3% (5.9%–7.0%)		10.3% (7.6%–13.1%)
HR (bin=5)	CTGAN CORE	12.5% ± 1.1%	14.1% ± 1.8%	15.2% ± 1.9%	16.3% ± 2.0%	18.1% ± 2.2%	19.6% ± 2.6%	20.8% ± 2.9%	22.6% ± 3.4%	24.2% ± 3.5%	26.1% ± 4.0%	27.0% ± 3.9%	32.7% ± 4.8%	37.8% ± 4.8%
HR* (bin=5)	CTGAN CORE	7.0% ± 0.6%	7.9% ± 1.0%	8.5% ± 1.0%	9.1% ± 1.1%	10.1% ± 1.2%	11.0% ± 1.5%	11.6% ± 1.6%	12.6% ± 1.9%	13.5% ± 2.0%	14.6% ± 2.2%	15.1% ± 2.2%	18.3% ± 2.7%	21.1% ± 2.7%
HR (bin=7)	CTGAN CORE	26.9% ± 1.5%	29.7% ± 2.2%	31.8% ± 2.5%	33.6% ± 2.5%	36.8% ± 2.9%	39.5% ± 3.2%	42.2% ± 3.9%	44.9% ± 4.2%	46.7% ± 4.1%	48.8% ± 4.5%	50.0% ± 4.7%	57.6% ± 5.3%	61.7% ± 5.2%
HR* (bin=7)	CTGAN CORE	13.2% ± 0.7%	14.6% ± 1.1%	15.6% ± 1.2%	16.5% ± 1.4%	18.0% ± 1.6%	19.3% ± 1.9%	20.7% ± 2.1%	22.0% ± 2.2%	22.9% ± 2.0%	23.9% ± 2.3%	24.5% ± 2.3%	28.2% ± 2.6%	30.2% ± 2.6%
HR (bin=10)	CTGAN CORE	50.0% ± 1.8%	53.3% ± 2.7%	55.6% ± 3.2%	58.1% ± 3.1%	61.3% ± 3.5%	63.6% ± 3.9%	65.4% ± 4.2%	67.2% ± 4.2%	68.9% ± 4.1%	70.9% ± 4.1%	71.6% ± 4.2%	77.5% ± 4.4%	80.9% ± 3.6%

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR* (bin=10)	CTGAN CORE	21.7% ± 0.8%	23.1% ± 1.2%	24.1% ± 1.4%	25.2% ± 1.4%	26.6% ± 1.5%	27.6% ± 1.7%	28.4% ± 1.8%	29.1% ± 1.8%	29.9% ± 1.8%	30.8% ± 1.8%	31.1% ± 1.8%	33.6% ± 1.9%	35.1% ± 1.6%
HR (bin=15)	CTGAN CORE	80.8% ± 1.1%	83.2% ± 1.6%	84.5% ± 1.6%	86.2% ± 1.5%	87.5% ± 1.7%	88.7% ± 1.7%	89.8% ± 1.8%	90.6% ± 1.6%	91.3% ± 1.7%	92.0% ± 1.9%	92.4% ± 1.8%	94.3% ± 1.9%	95.4% ± 1.3%
HR* (bin=15)	CTGAN CORE	31.0% ± 0.4%	31.9% ± 0.6%	32.4% ± 0.6%	33.1% ± 0.6%	33.6% ± 0.6%	34.0% ± 0.7%	34.4% ± 0.7%	34.8% ± 0.6%	35.0% ± 0.6%	35.3% ± 0.7%	35.4% ± 0.7%	36.2% ± 0.7%	36.6% ± 0.5%
HR (bin=20)	CTGAN CORE	94.6% ± 0.4%	95.5% ± 0.5%	95.9% ± 0.5%	96.4% ± 0.4%	97.0% ± 0.5%	97.3% ± 0.4%	97.5% ± 0.5%	97.8% ± 0.5%	97.8% ± 0.5%	98.1% ± 0.4%	98.2% ± 0.4%	98.8% ± 0.3%	99.1% ± 0.3%
HR* (bin=20)	CTGAN CORE	33.5% ± 0.1%	33.8% ± 0.2%	34.0% ± 0.2%	34.2% ± 0.1%	34.4% ± 0.2%	34.5% ± 0.1%	34.6% ± 0.2%	34.7% ± 0.2%	34.7% ± 0.2%	34.8% ± 0.2%	34.8% ± 0.2%	35.0% ± 0.1%	35.1% ± 0.1%
HR (bin=30)	CTGAN CORE	99.6% ± 0.0%	99.7% ± 0.0%	99.7% ± 0.0%	99.7% ± 0.0%	99.8% ± 0.1%	99.8% ± 0.0%	99.8% ± 0.0%	99.8% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	99.9% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	CTGAN CORE	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%
Expert rules (all) %	CTGAN CORE	1.0% ± 0.9%	0.7% ± 0.3%	0.7% ± 0.2%	0.9% ± 0.4%	1.0% ± 0.5%	1.2% ± 0.5%	1.4% ± 0.6%	1.6% ± 0.7%	1.9% ± 0.7%	1.9% ± 0.9%	2.1% ± 1.2%	2.2% ± 1.5%	2.7% ± 1.5%
% impossible expert rules	CTGAN CORE	31.60%	68.70%	85.40%	93.70%	94.10%	94.00%	88.80%	89.50%	91.10%	96.30%	95.70%	93.50%	93.80%
LLM only %	CTGAN CORE	2.1% (1.3%–3.2%)						5.5% (2.7%–5.6%)				8.0% (3.6%–8.1%)		5.8% (1.4%– 8.4%)
Expert rules + LLM %	CTGAN CORE	2.4% (1.6%–3.7%)						6.3% (4.2%–7.7%)				11.9% (4.4%–14.6%)		13.0% (3.6%– 15.7%)
HR (bin=5)	DDPM	89.8% ± 0.3%	92.5% ± 0.3%	95.4% ± 0.4%	98.3% ± 0.3%	99.7% ± 0.2%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=5)	DDPM	26.6% ± 0.1%	27.4% ± 0.1%	28.3% ± 0.1%	29.1% ± 0.1%	29.5% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%	29.6% ± 0.0%
HR (bin=7)	DDPM	97.6% ± 0.1%	98.6% ± 0.0%	99.2% ± 0.1%	99.7% ± 0.1%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=7)	DDPM	25.8% ± 0.0%	26.1% ± 0.0%	26.2% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%	26.4% ± 0.0%
HR (bin=10)	DDPM	99.6% ± 0.0%	99.8% ± 0.0%	99.9% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=10)	DDPM	23.6% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%
HR (bin=15)	DDPM	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=15)	DDPM	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%
HR (bin=20)	DDPM	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=20)	DDPM	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR (bin=30)	DDPM	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	DDPM	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%
Expert rules (all) %	DDPM	4.9% ± 0.7%	21.3% ± 2.3%	46.9% ± 3.6%	77.2% ± 3.5%	93.5% ± 2.9%	99.6% ± 0.6%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
% impossible expert rules	DDPM	32.10%	32.00%	35.50%	38.00%	39.30%	38.90%	39.30%	39.50%	39.70%	39.60%	39.50%	39.40%	39.40%
LLM only %	DDPM	3.5% (3.2%–4.2%)					0.0% (0.0%–0.1%)					0.0% (0.0%–0.0%)		0.0% (0.0%–0.0%)
Expert rules + LLM %	DDPM	6.9% (6.7%–8.7%)					100.0% (99.8%–100.0%)					100.0% (100.0%–100.0%)		100.0% (100.0%–100.0%)
HR (bin=5)	DDPM CORE	11.7% ± 0.2%	13.4% ± 0.4%	17.4% ± 1.0%	24.9% ± 1.7%	36.8% ± 2.9%	49.4% ± 3.5%	61.9% ± 3.6%	71.0% ± 2.8%	78.2% ± 2.3%	84.2% ± 2.1%	89.2% ± 2.1%	99.0% ± 0.8%	99.1% ± 0.7%
HR* (bin=5)	DDPM CORE	6.5% ± 0.1%	7.5% ± 0.2%	9.7% ± 0.6%	13.9% ± 1.0%	20.5% ± 1.6%	27.6% ± 2.0%	34.6% ± 2.0%	39.7% ± 1.6%	43.7% ± 1.3%	47.0% ± 1.2%	49.8% ± 1.2%	55.3% ± 0.5%	55.4% ± 0.4%
HR (bin=7)	DDPM CORE	25.0% ± 0.2%	26.6% ± 0.4%	30.4% ± 0.9%	37.2% ± 1.5%	48.1% ± 2.5%	59.6% ± 2.9%	70.8% ± 2.9%	78.8% ± 2.3%	84.9% ± 1.7%	89.8% ± 1.5%	93.7% ± 1.5%	99.8% ± 0.4%	99.8% ± 0.3%
HR* (bin=7)	DDPM CORE	12.3% ± 0.1%	13.1% ± 0.2%	14.9% ± 0.4%	18.3% ± 0.7%	23.6% ± 1.2%	29.2% ± 1.4%	34.7% ± 1.4%	38.7% ± 1.1%	41.6% ± 0.8%	44.0% ± 0.8%	46.0% ± 0.7%	48.9% ± 0.2%	49.0% ± 0.1%
HR (bin=10)	DDPM CORE	48.1% ± 0.2%	49.0% ± 0.3%	51.3% ± 0.6%	55.5% ± 1.1%	62.6% ± 1.8%	70.8% ± 2.2%	79.2% ± 2.2%	85.6% ± 1.7%	90.3% ± 1.2%	93.8% ± 1.0%	96.6% ± 0.9%	99.9% ± 0.1%	100.0% ± 0.1%
HR* (bin=10)	DDPM CORE	20.8% ± 0.1%	21.3% ± 0.1%	22.2% ± 0.2%	24.1% ± 0.5%	27.2% ± 0.8%	30.7% ± 1.0%	34.4% ± 1.0%	37.1% ± 0.7%	39.2% ± 0.5%	40.7% ± 0.4%	41.9% ± 0.4%	43.3% ± 0.0%	43.4% ± 0.0%
HR (bin=15)	DDPM CORE	79.6% ± 0.2%	80.3% ± 0.2%	80.8% ± 0.3%	81.7% ± 0.5%	83.4% ± 0.7%	85.7% ± 0.9%	88.8% ± 1.2%	91.7% ± 1.0%	94.2% ± 0.7%	96.2% ± 0.7%	97.8% ± 0.6%	100.0% ± 0.1%	100.0% ± 0.0%
HR* (bin=15)	DDPM CORE	30.6% ± 0.1%	30.8% ± 0.1%	31.0% ± 0.1%	31.4% ± 0.2%	32.0% ± 0.3%	32.9% ± 0.4%	34.1% ± 0.4%	35.2% ± 0.4%	36.1% ± 0.3%	36.9% ± 0.3%	37.5% ± 0.2%	38.3% ± 0.0%	38.3% ± 0.0%
HR (bin=20)	DDPM CORE	94.2% ± 0.2%	94.6% ± 0.1%	94.8% ± 0.1%	95.0% ± 0.2%	95.3% ± 0.2%	95.8% ± 0.2%	96.5% ± 0.3%	97.1% ± 0.3%	97.8% ± 0.3%	98.5% ± 0.3%	99.2% ± 0.2%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=20)	DDPM CORE	33.4% ± 0.1%	33.5% ± 0.0%	33.6% ± 0.1%	33.7% ± 0.1%	33.8% ± 0.1%	34.0% ± 0.1%	34.2% ± 0.1%	34.4% ± 0.1%	34.7% ± 0.1%	34.9% ± 0.1%	35.2% ± 0.1%	35.4% ± 0.0%	35.4% ± 0.0%
HR (bin=30)	DDPM CORE	99.5% ± 0.0%	99.6% ± 0.0%	99.6% ± 0.0%	99.6% ± 0.0%	99.6% ± 0.0%	99.6% ± 0.0%	99.7% ± 0.1%	99.8% ± 0.0%	99.8% ± 0.0%	99.8% ± 0.0%	99.9% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	DDPM CORE	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	31.9% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%	32.0% ± 0.0%
Expert rules (all) %	DDPM CORE	0.9% ± 0.1%	2.8% ± 0.3%	6.7% ± 0.9%	13.8% ± 1.7%	25.0% ± 2.9%	36.2% ± 3.7%	47.0% ± 4.3%	55.2% ± 4.4%	62.0% ± 3.9%	68.0% ± 4.1%	73.8% ± 4.0%	89.7% ± 3.7%	90.5% ± 3.6%
% impossible expert rules	DDPM CORE	25.10%	28.00%	32.60%	34.20%	33.50%	33.50%	33.90%	35.00%	35.10%	35.90%	35.80%	34.90%	33.80%
LLM only %	DDPM CORE	2.5% (2.2%–2.7%)					7.7% (4.9%–8.2%)					4.3% (1.1%–4.9%)		0.3% (0.0%–4.2%)

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Expert rules + LLM %	DDPM CORE	3.3% (2.9%–3.9%)					40.2% (38.8%–46.1%)					77.4% (75.6%–87.7%)		89.9% (88.5%–98.3%)
HR (bin=5)	RTVAE	85.8% ± 4.4%	86.8% ± 4.0%	88.7% ± 3.6%	89.6% ± 3.4%	91.0% ± 3.4%	90.7% ± 4.6%	90.6% ± 4.4%	91.8% ± 4.0%	93.1% ± 3.4%	94.3% ± 3.2%	94.1% ± 4.2%	96.3% ± 3.3%	97.0% ± 3.2%
HR* (bin=5)	RTVAE	25.4% ± 1.3%	25.7% ± 1.2%	26.3% ± 1.1%	26.6% ± 1.0%	26.9% ± 1.0%	26.9% ± 1.4%	26.8% ± 1.3%	27.2% ± 1.2%	27.6% ± 1.0%	27.9% ± 0.9%	27.9% ± 1.2%	28.5% ± 1.0%	28.7% ± 0.9%
HR (bin=7)	RTVAE	96.8% ± 1.2%	97.6% ± 0.9%	98.3% ± 0.7%	98.5% ± 0.7%	98.9% ± 0.5%	98.9% ± 0.6%	98.7% ± 0.7%	99.0% ± 0.6%	99.2% ± 0.5%	99.4% ± 0.5%	99.2% ± 0.9%	99.5% ± 0.8%	99.5% ± 0.8%
HR* (bin=7)	RTVAE	25.6% ± 0.3%	25.8% ± 0.2%	26.0% ± 0.2%	26.1% ± 0.2%	26.1% ± 0.1%	26.1% ± 0.2%	26.1% ± 0.2%	26.2% ± 0.2%	26.2% ± 0.1%	26.3% ± 0.1%	26.2% ± 0.2%	26.3% ± 0.2%	26.3% ± 0.2%
HR (bin=10)	RTVAE	99.5% ± 0.2%	99.7% ± 0.1%	99.8% ± 0.1%	99.8% ± 0.1%	99.9% ± 0.1%	99.9% ± 0.1%	99.8% ± 0.1%	99.9% ± 0.1%	99.9% ± 0.1%	99.9% ± 0.1%	99.9% ± 0.1%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=10)	RTVAE	23.6% ± 0.1%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%	23.7% ± 0.0%
HR (bin=15)	RTVAE	99.9% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=15)	RTVAE	21.2% ± 0.0%	21.2% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%	21.3% ± 0.0%
HR (bin=20)	RTVAE	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=20)	RTVAE	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%	19.8% ± 0.0%
HR (bin=30)	RTVAE	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%	100.0% ± 0.0%
HR* (bin=30)	RTVAE	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%	18.0% ± 0.0%
Expert rules (all) %	RTVAE	0.9% ± 0.9%	2.1% ± 3.3%	2.0% ± 3.4%	1.5% ± 2.6%	1.5% ± 2.7%	1.5% ± 2.7%	1.5% ± 2.8%	0.9% ± 1.8%	0.6% ± 1.2%	0.4% ± 0.9%	0.4% ± 0.9%	0.0% ± 0.0%	0.0% ± 0.0%
% impossible expert rules	RTVAE	65.80%	86.10%	90.40%	93.10%	96.00%	97.30%	97.20%	95.70%	96.60%	97.00%	97.80%	NA	NA
LLM only %	RTVAE	2.8% (0.8%–3.0%)					1.3% (0.0%–1.6%)					0.0% (0.0%–0.3%)		0.0% (0.0%–0.0%)
Expert rules + LLM %	RTVAE	6.6% (0.8%–7.9%)					3.3% (0.0%–20.6%)					0.0% (0.0%–6.6%)		0.0% (0.0%–0.0%)
HR (bin=5)	RTVAE CORE	13.7% ± 10.3%	12.6% ± 10.5%	11.8% ± 10.8%	11.5% ± 10.9%	11.4% ± 10.8%	11.9% ± 11.4%	12.1% ± 11.8%	12.0% ± 11.8%	11.9% ± 11.8%	11.9% ± 11.9%	11.8% ± 11.9%	11.8% ± 12.1%	11.6% ± 11.7%
HR* (bin=5)	RTVAE CORE	7.7% ± 5.7%	7.1% ± 5.9%	6.6% ± 6.0%	6.4% ± 6.1%	6.4% ± 6.1%	6.6% ± 6.3%	6.8% ± 6.6%	6.7% ± 6.6%	6.7% ± 6.6%	6.7% ± 6.6%	6.6% ± 6.6%	6.6% ± 6.7%	6.5% ± 6.6%
HR (bin=7)	RTVAE CORE	28.6% ± 14.6%	27.3% ± 15.1%	26.5% ± 15.1%	26.1% ± 14.9%	26.8% ± 15.2%	26.9% ± 15.6%	26.5% ± 16.0%	26.3% ± 15.9%	26.5% ± 16.0%	26.5% ± 16.9%	26.1% ± 17.0%	25.7% ± 17.5%	22.1% ± 18.0%
HR* (bin=7)	RTVAE CORE	14.0% ± 7.2%	13.4% ± 7.4%	13.0% ± 7.4%	12.8% ± 7.3%	13.1% ± 7.5%	13.2% ± 7.6%	13.0% ± 7.9%	12.9% ± 7.8%	13.0% ± 7.9%	13.0% ± 8.3%	12.8% ± 8.4%	12.6% ± 8.6%	10.8% ± 8.8%

metric	Gen.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
HR (bin=10)	RTVAE CORE	46.4% ± 15.5%	44.3% ± 16.3%	43.0% ± 16.6%	42.7% ± 16.7%	43.2% ± 16.8%	42.8% ± 17.2%	41.8% ± 17.9%	41.4% ± 17.9%	41.7% ± 17.9%	41.0% ± 18.4%	40.2% ± 18.8%	38.0% ± 19.3%	32.6% ± 20.7%
HR* (bin=10)	RTVAE CORE	20.1% ± 6.7%	19.2% ± 7.1%	18.6% ± 7.2%	18.5% ± 7.2%	18.7% ± 7.3%	18.6% ± 7.4%	18.1% ± 7.8%	18.0% ± 7.8%	18.1% ± 7.8%	17.8% ± 8.0%	17.5% ± 8.2%	16.5% ± 8.4%	14.1% ± 9.0%
HR (bin=15)	RTVAE CORE	72.5% ± 9.2%	70.1% ± 10.2%	68.9% ± 10.7%	68.4% ± 10.9%	68.5% ± 11.1%	67.7% ± 11.7%	66.3% ± 12.7%	66.4% ± 12.8%	66.6% ± 12.9%	65.7% ± 13.7%	64.7% ± 14.0%	61.4% ± 15.5%	55.1% ± 17.0%
HR* (bin=15)	RTVAE CORE	27.8% ± 3.5%	26.9% ± 3.9%	26.4% ± 4.1%	26.3% ± 4.2%	26.3% ± 4.3%	26.0% ± 4.5%	25.4% ± 4.9%	25.5% ± 4.9%	25.5% ± 5.0%	25.2% ± 5.3%	24.8% ± 5.4%	23.6% ± 6.0%	21.1% ± 6.5%
HR (bin=20)	RTVAE CORE	92.8% ± 2.5%	92.1% ± 2.8%	91.8% ± 2.9%	91.7% ± 3.0%	91.7% ± 3.0%	91.5% ± 3.2%	90.8% ± 3.7%	90.9% ± 3.6%	91.0% ± 3.6%	90.7% ± 3.8%	90.3% ± 4.0%	89.2% ± 4.5%	87.5% ± 4.9%
HR* (bin=20)	RTVAE CORE	32.9% ± 0.9%	32.7% ± 1.0%	32.6% ± 1.0%	32.5% ± 1.0%	32.5% ± 1.1%	32.4% ± 1.1%	32.2% ± 1.3%	32.2% ± 1.3%	32.3% ± 1.3%	32.2% ± 1.4%	32.0% ± 1.4%	31.6% ± 1.6%	31.0% ± 1.7%
HR (bin=30)	RTVAE CORE	99.1% ± 0.3%	99.1% ± 0.3%	99.0% ± 0.4%	99.0% ± 0.4%	99.0% ± 0.4%	98.8% ± 0.5%	98.8% ± 0.5%	98.8% ± 0.6%	98.8% ± 0.5%	98.6% ± 0.6%	98.6% ± 0.6%	98.3% ± 0.8%	98.0% ± 0.9%
HR* (bin=30)	RTVAE CORE	31.7% ± 0.1%	31.7% ± 0.1%	31.7% ± 0.1%	31.7% ± 0.1%	31.7% ± 0.1%	31.7% ± 0.2%	31.6% ± 0.2%	31.6% ± 0.2%	31.6% ± 0.2%	31.6% ± 0.2%	31.6% ± 0.2%	31.5% ± 0.3%	31.4% ± 0.3%
Expert rules (all) %	RTVAE CORE	9.8% ± 13.5%	10.8% ± 13.5%	10.5% ± 13.2%	10.2% ± 12.9%	10.2% ± 13.1%	9.3% ± 12.4%	9.3% ± 12.5%	9.2% ± 12.5%	9.0% ± 12.6%	8.6% ± 12.5%	8.5% ± 12.2%	8.4% ± 12.5%	7.8% ± 12.6%
% impossible expert rules	RTVAE CORE	36.90%	40.90%	40.60%	38.10%	37.60%	35.30%	34.70%	33.80%	31.20%	28.30%	28.70%	23.50%	17.70%
LLM only %	RTVAE CORE	0.4% (0.2%–0.6%)					0.0% (0.0%–0.8%)					0.0% (0.0%–0.8%)		0.0% (0.0%–0.0%)
Expert rules + LLM %	RTVAE CORE	0.5% (0.2%–0.6%)					0.0% (0.0%–0.8%)					0.0% (0.0%–0.8%)		0.0% (0.0%–0.0%)

Table 2 Quality Evaluation results for all generators

Results from full and low complexity (CORE) dataset. Reported values are the means $\pm$ half-width of 95 % CI across 15 chains of recursive generations from the different data seed – model seed combinations for all generators and the two versions of the dataset – full variables and core variables.

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
alpha-Precision	ARF	0.9336 $\pm$ 0.0061	0.8749 $\pm$ 0.0074	0.8312 $\pm$ 0.0085	0.7996 $\pm$ 0.0071	0.7757 $\pm$ 0.0089	0.7545 $\pm$ 0.0097	0.7418 $\pm$ 0.0093	0.7319 $\pm$ 0.0089	0.7189 $\pm$ 0.0122	0.7146 $\pm$ 0.0156	0.7052 $\pm$ 0.0139	0.6778 $\pm$ 0.0220	0.6614 $\pm$ 0.0267
beta-Recall	ARF	0.4409 $\pm$ 0.0068	0.3946 $\pm$ 0.0067	0.3562 $\pm$ 0.0096	0.3301 $\pm$ 0.0104	0.3092 $\pm$ 0.0135	0.2920 $\pm$ 0.0103	0.2771 $\pm$ 0.0148	0.2642 $\pm$ 0.0130	0.2522 $\pm$ 0.0125	0.2431 $\pm$ 0.0131	0.2352 $\pm$ 0.0143	0.2077 $\pm$ 0.0137	0.1959 $\pm$ 0.0167
Authenticity	ARF	0.5255 $\pm$ 0.0054	0.5515 $\pm$ 0.0054	0.5736 $\pm$ 0.0092	0.5992 $\pm$ 0.0080	0.6143 $\pm$ 0.0127	0.6312 $\pm$ 0.0139	0.6463 $\pm$ 0.0141	0.6609 $\pm$ 0.0151	0.6724 $\pm$ 0.0134	0.6813 $\pm$ 0.0106	0.6886 $\pm$ 0.0130	0.7170 $\pm$ 0.0173	0.7335 $\pm$ 0.0167
PRDC Precision	ARF	0.9262 $\pm$ 0.0039	0.8943 $\pm$ 0.0045	0.8732 $\pm$ 0.0076	0.8551 $\pm$ 0.0059	0.8451 $\pm$ 0.0067	0.8363 $\pm$ 0.0064	0.8315 $\pm$ 0.0083	0.8284 $\pm$ 0.0100	0.8273 $\pm$ 0.0094	0.8251 $\pm$ 0.0090	0.8239 $\pm$ 0.0102	0.8214 $\pm$ 0.0136	0.8214 $\pm$ 0.0168
PRDC Recall	ARF	0.9736 $\pm$ 0.0022	0.9698 $\pm$ 0.0026	0.9682 $\pm$ 0.0023	0.9650 $\pm$ 0.0021	0.9623 $\pm$ 0.0037	0.9591 $\pm$ 0.0037	0.9557 $\pm$ 0.0044	0.9519 $\pm$ 0.0059	0.9482 $\pm$ 0.0069	0.9431 $\pm$ 0.0069	0.9406 $\pm$ 0.0063	0.9269 $\pm$ 0.0129	0.9093 $\pm$ 0.0129
PRDC Density	ARF	0.8307 $\pm$ 0.0118	0.7255 $\pm$ 0.0119	0.6633 $\pm$ 0.0145	0.6256 $\pm$ 0.0142	0.6021 $\pm$ 0.0137	0.5873 $\pm$ 0.0132	0.5757 $\pm$ 0.0170	0.5673 $\pm$ 0.0164	0.5642 $\pm$ 0.0182	0.5601 $\pm$ 0.0209	0.5577 $\pm$ 0.0212	0.5493 $\pm$ 0.0278	0.5441 $\pm$ 0.0330
PRDC Coverage	ARF	0.9296 $\pm$ 0.0047	0.8638 $\pm$ 0.0061	0.7965 $\pm$ 0.0140	0.7366 $\pm$ 0.0152	0.6857 $\pm$ 0.0150	0.6416 $\pm$ 0.0184	0.5996 $\pm$ 0.0199	0.5620 $\pm$ 0.0214	0.5303 $\pm$ 0.0191	0.5049 $\pm$ 0.0185	0.4813 $\pm$ 0.0191	0.4146 $\pm$ 0.0146	0.3757 $\pm$ 0.0179
TV Complement	ARF	0.9749 $\pm$ 0.0018	0.9623 $\pm$ 0.0018	0.9548 $\pm$ 0.0023	0.9499 $\pm$ 0.0027	0.9462 $\pm$ 0.0028	0.9431 $\pm$ 0.0028	0.9409 $\pm$ 0.0033	0.9390 $\pm$ 0.0033	0.9371 $\pm$ 0.0035	0.9360 $\pm$ 0.0033	0.9351 $\pm$ 0.0037	0.9296 $\pm$ 0.0039	0.9272 $\pm$ 0.0041
KS Complement	ARF	0.9464 $\pm$ 0.0038	0.9135 $\pm$ 0.0034	0.8872 $\pm$ 0.0047	0.8645 $\pm$ 0.0047	0.8445 $\pm$ 0.0047	0.8260 $\pm$ 0.0048	0.8097 $\pm$ 0.0049	0.7953 $\pm$ 0.0057	0.7826 $\pm$ 0.0050	0.7717 $\pm$ 0.0051	0.7632 $\pm$ 0.0055	0.7302 $\pm$ 0.0059	0.7138 $\pm$ 0.0091
Wasserstein Distance	ARF	0.1078 $\pm$ 0.0025	0.1170 $\pm$ 0.0029	0.1232 $\pm$ 0.0030	0.1278 $\pm$ 0.0031	0.1310 $\pm$ 0.0035	0.1338 $\pm$ 0.0037	0.1361 $\pm$ 0.0042	0.1376 $\pm$ 0.0042	0.1393 $\pm$ 0.0047	0.1409 $\pm$ 0.0049	0.1420 $\pm$ 0.0050	0.1473 $\pm$ 0.0065	0.1506 $\pm$ 0.0070
Maximum Mean Discrepancy	ARF	0.0023 $\pm$ 0.0002	0.0073 $\pm$ 0.0006	0.0134 $\pm$ 0.0011	0.0198 $\pm$ 0.0016	0.0267 $\pm$ 0.0021	0.0337 $\pm$ 0.0027	0.0407 $\pm$ 0.0031	0.0475 $\pm$ 0.0036	0.0539 $\pm$ 0.0039	0.0597 $\pm$ 0.0042	0.0651 $\pm$ 0.0046	0.0871 $\pm$ 0.0062	0.1008 $\pm$ 0.0079
Jensen-Shannon Divergence	ARF	0.0660 $\pm$ 0.0013	0.0914 $\pm$ 0.0012	0.1116 $\pm$ 0.0020	0.1267 $\pm$ 0.0023	0.1384 $\pm$ 0.0023	0.1489 $\pm$ 0.0028	0.1579 $\pm$ 0.0035	0.1659 $\pm$ 0.0036	0.1727 $\pm$ 0.0036	0.1787 $\pm$ 0.0034	0.1837 $\pm$ 0.0034	0.2021 $\pm$ 0.0034	0.2112 $\pm$ 0.0043
Detector Average	ARF	0.6340 $\pm$ 0.0089	0.7141 $\pm$ 0.0103	0.7628 $\pm$ 0.0110	0.7852 $\pm$ 0.0159	0.8015 $\pm$ 0.0214	0.8174 $\pm$ 0.0178	0.8388 $\pm$ 0.0199	0.8418 $\pm$ 0.0277	0.8602 $\pm$ 0.0189	0.8666 $\pm$ 0.0305	0.8669 $\pm$ 0.0331	0.8785 $\pm$ 0.0328	0.8950 $\pm$ 0.0260
GMM Detector	ARF	0.5607 $\pm$ 0.0220	0.5986 $\pm$ 0.0281	0.6405 $\pm$ 0.0342	0.6290 $\pm$ 0.0635	0.6247 $\pm$ 0.0782	0.6237 $\pm$ 0.0697	0.6563 $\pm$ 0.0894	0.6293 $\pm$ 0.1026	0.6693 $\pm$ 0.0647	0.6691 $\pm$ 0.1138	0.6547 $\pm$ 0.1264	0.6471 $\pm$ 0.1252	0.6967 $\pm$ 0.0998
XGBoost Detector	ARF	0.7818 $\pm$ 0.0084	0.8860 $\pm$ 0.0060	0.9270 $\pm$ 0.0034	0.9481 $\pm$ 0.0036	0.9598 $\pm$ 0.0019	0.9681 $\pm$ 0.0019	0.9747 $\pm$ 0.0021	0.9793 $\pm$ 0.0016	0.9829 $\pm$ 0.0014	0.9853 $\pm$ 0.0011	0.9871 $\pm$ 0.0011	0.9918 $\pm$ 0.0013	0.9934 $\pm$ 0.0011
MLP Detector	ARF	0.5785 $\pm$ 0.0240	0.6813 $\pm$ 0.0139	0.7362 $\pm$ 0.0112	0.7761 $\pm$ 0.0166	0.8052 $\pm$ 0.0205	0.8351 $\pm$ 0.0181	0.8612 $\pm$ 0.0151	0.8777 $\pm$ 0.0176	0.8944 $\pm$ 0.0094	0.9072 $\pm$ 0.0102	0.9128 $\pm$ 0.0119	0.9378 $\pm$ 0.0106	0.9427 $\pm$ 0.0097
Linear Detector	ARF	0.6151 $\pm$ 0.0075	0.6904 $\pm$ 0.0078	0.7477 $\pm$ 0.0088	0.7877 $\pm$ 0.0099	0.8163 $\pm$ 0.0101	0.8428 $\pm$ 0.0084	0.8631 $\pm$ 0.0083	0.8811 $\pm$ 0.0088	0.8942 $\pm$ 0.0080	0.9046 $\pm$ 0.0074	0.9128 $\pm$ 0.0074	0.9375 $\pm$ 0.0066	0.9472 $\pm$ 0.0074

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
SDMetrics Column Shapes	ARF	0.9600 ± 0.0019	0.9368 ± 0.0016	0.9195 ± 0.0022	0.9053 ± 0.0023	0.8931 ± 0.0022	0.8820 ± 0.0025	0.8725 ± 0.0029	0.8640 ± 0.0031	0.8565 ± 0.0029	0.8503 ± 0.0030	0.8454 ± 0.0031	0.8256 ± 0.0038	0.8159 ± 0.0050
SDMetrics Column Pair Trends	ARF	0.8515 ± 0.0095	0.7886 ± 0.0102	0.7537 ± 0.0126	0.7305 ± 0.0147	0.7143 ± 0.0168	0.7007 ± 0.0173	0.6891 ± 0.0180	0.6804 ± 0.0196	0.6737 ± 0.0203	0.6665 ± 0.0208	0.6621 ± 0.0222	0.6572 ± 0.0229	0.6542 ± 0.0240
Memorized Fraction	ARF	0.0020 ± 0.0062	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
New Row Synthesis	ARF	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Generation Time (s)	ARF	11.1833 ± 0.9124	11.3547 ± 0.7826	11.7120 ± 0.9740	11.6693 ± 0.3229	11.7360 ± 0.3994	11.7560 ± 0.4038	11.7273 ± 0.4503	11.9220 ± 0.4982	11.6327 ± 0.3884	11.7327 ± 0.3608	11.7520 ± 0.4243	11.7447 ± 0.3909	11.6373 ± 0.3848
Model Size (MB)	ARF	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	ARF	0.1265 ± 0.0034	0.1245 ± 0.0067	0.1246 ± 0.0079	0.1247 ± 0.0106	0.1243 ± 0.0129	0.1235 ± 0.0147	0.1241 ± 0.0159	0.1239 ± 0.0180	0.1235 ± 0.0201	0.1238 ± 0.0230	0.1220 ± 0.0248	0.1221 ± 0.0314	0.1213 ± 0.0372
Readmission Collapsed	ARF	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TRTR AUROC	ARF	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054
TSTR AUROC	ARF	0.6204 ± 0.0097	0.6000 ± 0.0108	0.5737 ± 0.0513	0.5675 ± 0.0245	0.5435 ± 0.0389	0.5376 ± 0.0372	0.5300 ± 0.0396	0.5170 ± 0.0405	0.5112 ± 0.0172	0.5056 ± 0.0331	0.5075 ± 0.0431	0.4964 ± 0.0531	0.4936 ± 0.0354
TSTR Accuracy	ARF	0.6563 ± 0.0824	0.7619 ± 0.1045	0.7689 ± 0.1272	0.7692 ± 0.1537	0.7925 ± 0.1316	0.7349 ± 0.2138	0.7332 ± 0.1741	0.7297 ± 0.2685	0.7050 ± 0.2939	0.6606 ± 0.2212	0.6780 ± 0.2605	0.6031 ± 0.3127	0.5285 ± 0.3331
TSTR Precision	ARF	0.1817 ± 0.0103	0.1577 ± 0.0719	0.1409 ± 0.0833	0.1016 ± 0.0969	0.0922 ± 0.0899	0.1052 ± 0.0950	0.1042 ± 0.0808	0.0809 ± 0.0802	0.0602 ± 0.0745	0.0968 ± 0.0686	0.0880 ± 0.0739	0.0871 ± 0.0724	0.0935 ± 0.0650
TSTR Recall	ARF	0.4749 ± 0.1429	0.2398 ± 0.2136	0.2009 ± 0.2328	0.1899 ± 0.2623	0.1333 ± 0.2146	0.2106 ± 0.2873	0.2243 ± 0.2593	0.1985 ± 0.3638	0.2273 ± 0.3962	0.2933 ± 0.2938	0.2710 ± 0.3608	0.3759 ± 0.4179	0.4592 ± 0.4381
TSTR F1	ARF	0.2578 ± 0.0165	0.1685 ± 0.0979	0.1363 ± 0.1001	0.1106 ± 0.1197	0.0862 ± 0.1054	0.1116 ± 0.1094	0.1248 ± 0.1099	0.0837 ± 0.1040	0.0850 ± 0.1089	0.1348 ± 0.1005	0.1110 ± 0.1115	0.1278 ± 0.1198	0.1465 ± 0.1061
TSTR FPR	ARF	0.3173 ± 0.1151	0.1620 ± 0.1508	0.1483 ± 0.1797	0.1463 ± 0.2140	0.1114 ± 0.1820	0.1886 ± 0.2863	0.1926 ± 0.2367	0.1928 ± 0.3605	0.2253 ± 0.3945	0.2858 ± 0.2959	0.2627 ± 0.3506	0.3638 ± 0.4188	0.4614 ± 0.4455
TSTR FNR	ARF	0.5251 ± 0.1429	0.7602 ± 0.2136	0.7991 ± 0.2328	0.8101 ± 0.2623	0.8667 ± 0.2146	0.7894 ± 0.2873	0.7757 ± 0.2593	0.8015 ± 0.3638	0.7727 ± 0.3962	0.7067 ± 0.2938	0.7290 ± 0.3608	0.6241 ± 0.4179	0.5408 ± 0.4381
TSTR TNR	ARF	0.6827 ± 0.1151	0.8380 ± 0.1508	0.8517 ± 0.1797	0.8537 ± 0.2140	0.8886 ± 0.1820	0.8114 ± 0.2863	0.8074 ± 0.2367	0.8072 ± 0.3605	0.7747 ± 0.3945	0.7142 ± 0.2959	0.7373 ± 0.3506	0.6362 ± 0.4188	0.5386 ± 0.4455
alpha-Precision	ARF CORE	0.9690 ± 0.0068	0.9434 ± 0.0097	0.9237 ± 0.0110	0.9063 ± 0.0128	0.8923 ± 0.0145	0.8785 ± 0.0146	0.8708 ± 0.0126	0.8657 ± 0.0125	0.8603 ± 0.0110	0.8564 ± 0.0108	0.8534 ± 0.0095	0.8501 ± 0.0092	0.8519 ± 0.0105
beta-Recall	ARF CORE	0.4764 ± 0.0089	0.4451 ± 0.0072	0.4211 ± 0.0112	0.3952 ± 0.0104	0.3766 ± 0.0145	0.3564 ± 0.0142	0.3440 ± 0.0149	0.3299 ± 0.0146	0.3194 ± 0.0137	0.3069 ± 0.0134	0.2981 ± 0.0179	0.2660 ± 0.0205	0.2481 ± 0.0199
Authenticity	ARF CORE	0.5106 ± 0.0055	0.5291 ± 0.0057	0.5432 ± 0.0098	0.5566 ± 0.0076	0.5734 ± 0.0098	0.5864 ± 0.0120	0.5947 ± 0.0116	0.6066 ± 0.0138	0.6141 ± 0.0120	0.6303 ± 0.0142	0.6310 ± 0.0141	0.6684 ± 0.0196	0.6906 ± 0.0175

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
PRDC Precision	ARF CORE	0.9646 ± 0.0019	0.9453 ± 0.0027	0.9293 ± 0.0065	0.9144 ± 0.0073	0.9021 ± 0.0076	0.8936 ± 0.0097	0.8834 ± 0.0119	0.8767 ± 0.0116	0.8715 ± 0.0114	0.8669 ± 0.0097	0.8647 ± 0.0127	0.8610 ± 0.0127	0.8607 ± 0.0129
PRDC Recall	ARF CORE	0.9870 ± 0.0011	0.9851 ± 0.0016	0.9848 ± 0.0018	0.9837 ± 0.0013	0.9824 ± 0.0019	0.9816 ± 0.0018	0.9797 ± 0.0018	0.9782 ± 0.0018	0.9765 ± 0.0023	0.9753 ± 0.0024	0.9743 ± 0.0024	0.9642 ± 0.0040	0.9546 ± 0.0054
PRDC Density	ARF CORE	0.9280 ± 0.0052	0.8461 ± 0.0078	0.7956 ± 0.0098	0.7557 ± 0.0082	0.7233 ± 0.0108	0.7006 ± 0.0133	0.6831 ± 0.0139	0.6678 ± 0.0157	0.6575 ± 0.0156	0.6501 ± 0.0141	0.6461 ± 0.0147	0.6382 ± 0.0166	0.6400 ± 0.0147
PRDC Coverage	ARF CORE	0.9586 ± 0.0036	0.9305 ± 0.0049	0.9005 ± 0.0062	0.8697 ± 0.0104	0.8383 ± 0.0131	0.8075 ± 0.0164	0.7806 ± 0.0220	0.7502 ± 0.0188	0.7233 ± 0.0227	0.6991 ± 0.0228	0.6757 ± 0.0249	0.5712 ± 0.0328	0.5103 ± 0.0251
TV Complement	ARF CORE	0.9850 ± 0.0014	0.9735 ± 0.0019	0.9645 ± 0.0023	0.9587 ± 0.0023	0.9545 ± 0.0020	0.9507 ± 0.0029	0.9485 ± 0.0024	0.9467 ± 0.0019	0.9457 ± 0.0013	0.9447 ± 0.0018	0.9443 ± 0.0021	0.9424 ± 0.0022	0.9412 ± 0.0023
KS Complement	ARF CORE	0.9714 ± 0.0024	0.9487 ± 0.0042	0.9279 ± 0.0051	0.9102 ± 0.0070	0.8928 ± 0.0088	0.8774 ± 0.0099	0.8624 ± 0.0107	0.8486 ± 0.0105	0.8356 ± 0.0116	0.8241 ± 0.0119	0.8134 ± 0.0135	0.7669 ± 0.0178	0.7417 ± 0.0195
Wasserstein Distance	ARF CORE	0.0179 ± 0.0004	0.0205 ± 0.0004	0.0230 ± 0.0005	0.0250 ± 0.0008	0.0269 ± 0.0009	0.0288 ± 0.0011	0.0304 ± 0.0011	0.0318 ± 0.0011	0.0330 ± 0.0012	0.0342 ± 0.0012	0.0352 ± 0.0011	0.0396 ± 0.0016	0.0422 ± 0.0022
Maximum Mean Discrepancy	ARF CORE	0.0009 ± 0.0003	0.0034 ± 0.0007	0.0072 ± 0.0013	0.0115 ± 0.0021	0.0168 ± 0.0030	0.0227 ± 0.0039	0.0291 ± 0.0046	0.0358 ± 0.0051	0.0427 ± 0.0061	0.0495 ± 0.0064	0.0566 ± 0.0074	0.0920 ± 0.0118	0.1149 ± 0.0140
Jensen-Shannon Divergence	ARF CORE	0.0579 ± 0.0015	0.0764 ± 0.0024	0.0936 ± 0.0031	0.1075 ± 0.0042	0.1208 ± 0.0049	0.1323 ± 0.0054	0.1421 ± 0.0057	0.1509 ± 0.0056	0.1582 ± 0.0057	0.1650 ± 0.0056	0.1712 ± 0.0066	0.1970 ± 0.0079	0.2115 ± 0.0091
Detector Average	ARF CORE	0.5587 ± 0.0067	0.6206 ± 0.0052	0.6627 ± 0.0133	0.6988 ± 0.0131	0.7139 ± 0.0172	0.7454 ± 0.0178	0.7629 ± 0.0151	0.7813 ± 0.0286	0.7991 ± 0.0222	0.8144 ± 0.0223	0.8206 ± 0.0259	0.8532 ± 0.0301	0.8724 ± 0.0285
GMM Detector	ARF CORE	0.5356 ± 0.0215	0.5789 ± 0.0185	0.5948 ± 0.0443	0.6311 ± 0.0440	0.5999 ± 0.0795	0.6480 ± 0.0664	0.6520 ± 0.0615	0.6658 ± 0.1037	0.6850 ± 0.0816	0.7054 ± 0.0724	0.6873 ± 0.1015	0.6848 ± 0.1126	0.7097 ± 0.1204
XGBoost Detector	ARF CORE	0.6065 ± 0.0082	0.7127 ± 0.0073	0.7765 ± 0.0059	0.8193 ± 0.0064	0.8496 ± 0.0064	0.8740 ± 0.0063	0.8905 ± 0.0056	0.9044 ± 0.0066	0.9158 ± 0.0066	0.9247 ± 0.0052	0.9327 ± 0.0054	0.9586 ± 0.0043	0.9681 ± 0.0032
MLP Detector	ARF CORE	0.5487 ± 0.0099	0.6090 ± 0.0093	0.6607 ± 0.0094	0.6962 ± 0.0102	0.7286 ± 0.0104	0.7554 ± 0.0105	0.7769 ± 0.0114	0.7976 ± 0.0091	0.8170 ± 0.0104	0.8291 ± 0.0108	0.8457 ± 0.0097	0.8899 ± 0.0083	0.9095 ± 0.0073
Linear Detector	ARF CORE	0.5441 ± 0.0103	0.5819 ± 0.0118	0.6187 ± 0.0157	0.6486 ± 0.0160	0.6776 ± 0.0144	0.7044 ± 0.0153	0.7322 ± 0.0173	0.7575 ± 0.0135	0.7786 ± 0.0177	0.7983 ± 0.0171	0.8167 ± 0.0133	0.8795 ± 0.0123	0.9025 ± 0.0091
SDMetrics Column Shapes	ARF CORE	0.9766 ± 0.0016	0.9582 ± 0.0029	0.9419 ± 0.0038	0.9288 ± 0.0050	0.9165 ± 0.0058	0.9056 ± 0.0066	0.8955 ± 0.0071	0.8863 ± 0.0068	0.8779 ± 0.0073	0.8705 ± 0.0074	0.8637 ± 0.0083	0.8344 ± 0.0111	0.8184 ± 0.0125
SDMetrics Column Pair Trends	ARF CORE	0.9455 ± 0.0049	0.9033 ± 0.0052	0.8730 ± 0.0064	0.8486 ± 0.0062	0.8293 ± 0.0062	0.8122 ± 0.0073	0.7970 ± 0.0076	0.7850 ± 0.0069	0.7720 ± 0.0060	0.7606 ± 0.0056	0.7497 ± 0.0067	0.7047 ± 0.0088	0.6870 ± 0.0053
Memorized Fraction	ARF CORE	1.3633 ± 0.1151	0.2367 ± 0.0532	0.1373 ± 0.0249	0.0867 ± 0.0342	0.0447 ± 0.0274	0.0320 ± 0.0193	0.0187 ± 0.0161	0.0113 ± 0.0213	0.0093 ± 0.0129	0.0093 ± 0.0088	0.0060 ± 0.0070	0.0013 ± 0.0039	0.0000 ± 0.0000
New Row Synthesis	ARF CORE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	ARF CORE	15.7460 ± 0.6923	15.9793 ± 0.6928	16.1133 ± 0.6445	16.0147 ± 0.4165	16.2667 ± 0.4607	16.3467 ± 0.5645	16.3293 ± 0.4554	16.3140 ± 0.3983	16.3960 ± 0.3393	16.4313 ± 0.4627	16.4100 ± 0.5292	12.4307 ± 7.9927	0.9113 ± 0.0332

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Generation Time (s)	ARF CORE	5.7920 ± 0.6229	5.8453 ± 0.5803	5.9527 ± 0.5593	5.8367 ± 0.3178	5.9627 ± 0.5384	5.9433 ± 0.3875	5.8733 ± 0.2026	5.9180 ± 0.1882	5.8913 ± 0.1185	5.9740 ± 0.2577	6.0133 ± 0.2278	5.9840 ± 0.2858	6.0913 ± 0.3124
Model Size (MB)	ARF CORE	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	ARF CORE	0.1258 ± 0.0041	0.1236 ± 0.0058	0.1206 ± 0.0053	0.1193 ± 0.0046	0.1187 ± 0.0040	0.1174 ± 0.0044	0.1166 ± 0.0052	0.1156 ± 0.0067	0.1150 ± 0.0074	0.1147 ± 0.0077	0.1145 ± 0.0081	0.1108 ± 0.0105	0.1085 ± 0.0144
Readmission Collapsed	ARF CORE	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TRTR AUROC	ARF CORE	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078
TSTR AUROC	ARF CORE	0.6049 ± 0.0093	0.6024 ± 0.0148	0.5888 ± 0.0162	0.5842 ± 0.0182	0.5821 ± 0.0155	0.5704 ± 0.0230	0.5543 ± 0.0474	0.5519 ± 0.0392	0.5422 ± 0.0521	0.5386 ± 0.0449	0.5451 ± 0.0394	0.5010 ± 0.0209	0.5144 ± 0.0315
TSTR Accuracy	ARF CORE	0.5913 ± 0.0829	0.6138 ± 0.0619	0.6732 ± 0.1123	0.7293 ± 0.1054	0.7187 ± 0.1081	0.7713 ± 0.0898	0.7291 ± 0.1666	0.7570 ± 0.2006	0.7563 ± 0.2013	0.7391 ± 0.2588	0.7832 ± 0.2075	0.8112 ± 0.2009	0.7450 ± 0.2873
TSTR Precision	ARF CORE	0.1706 ± 0.0128	0.1702 ± 0.0081	0.1599 ± 0.0503	0.1573 ± 0.0733	0.1756 ± 0.0208	0.1654 ± 0.0768	0.1619 ± 0.0561	0.1557 ± 0.0755	0.1440 ± 0.0760	0.0809 ± 0.0936	0.1174 ± 0.0973	0.0880 ± 0.0977	0.0510 ± 0.0869
TSTR Recall	ARF CORE	0.5602 ± 0.1085	0.5222 ± 0.1077	0.4026 ± 0.2021	0.3020 ± 0.2027	0.3184 ± 0.1795	0.2304 ± 0.1746	0.2612 ± 0.2059	0.2107 ± 0.2676	0.2026 ± 0.2743	0.1980 ± 0.3490	0.1569 ± 0.2836	0.0859 ± 0.2664	0.1681 ± 0.3740
TSTR F1	ARF CORE	0.2589 ± 0.0064	0.2548 ± 0.0162	0.2220 ± 0.0761	0.1897 ± 0.1034	0.2089 ± 0.0693	0.1787 ± 0.0900	0.1772 ± 0.0780	0.1485 ± 0.0929	0.1380 ± 0.0954	0.0954 ± 0.1107	0.1038 ± 0.1031	0.0408 ± 0.0771	0.0500 ± 0.0979
TSTR FPR	ARF CORE	0.4042 ± 0.1107	0.3728 ± 0.0862	0.2874 ± 0.1577	0.2083 ± 0.1499	0.2229 ± 0.1491	0.1498 ± 0.1281	0.2026 ± 0.2187	0.1633 ± 0.2680	0.1629 ± 0.2698	0.1820 ± 0.3470	0.1254 ± 0.2786	0.0830 ± 0.2690	0.1708 ± 0.3837
TSTR FNR	ARF CORE	0.4398 ± 0.1085	0.4778 ± 0.1077	0.5974 ± 0.2021	0.6980 ± 0.2027	0.6816 ± 0.1795	0.7696 ± 0.1746	0.7388 ± 0.2059	0.7893 ± 0.2676	0.7974 ± 0.2743	0.8020 ± 0.3490	0.8431 ± 0.2836	0.9141 ± 0.2664	0.8319 ± 0.3740
TSTR TNR	ARF CORE	0.5958 ± 0.1107	0.6272 ± 0.0862	0.7126 ± 0.1577	0.7917 ± 0.1499	0.7771 ± 0.1491	0.8502 ± 0.1281	0.7974 ± 0.2187	0.8367 ± 0.2680	0.8371 ± 0.2698	0.8180 ± 0.3470	0.8746 ± 0.2786	0.9170 ± 0.2690	0.8292 ± 0.3837
alpha-Precision	CTGAN	0.9702 ± 0.0207	0.9677 ± 0.0168	0.9665 ± 0.0169	0.9642 ± 0.0163	0.9596 ± 0.0148	0.9452 ± 0.0342	0.9271 ± 0.0711	0.9063 ± 0.0997	0.8977 ± 0.1085	0.8824 ± 0.1141	0.8739 ± 0.1210	0.8720 ± 0.1124	0.8544 ± 0.1149
beta-Recall	CTGAN	0.4776 ± 0.0238	0.4525 ± 0.0274	0.4443 ± 0.0298	0.4307 ± 0.0268	0.4179 ± 0.0270	0.4097 ± 0.0301	0.4000 ± 0.0329	0.3988 ± 0.0350	0.3905 ± 0.0361	0.3890 ± 0.0326	0.3776 ± 0.0384	0.3433 ± 0.0390	0.3313 ± 0.0444
Authenticity	CTGAN	0.5107 ± 0.0101	0.5204 ± 0.0174	0.5247 ± 0.0146	0.5317 ± 0.0134	0.5390 ± 0.0122	0.5507 ± 0.0176	0.5618 ± 0.0252	0.5713 ± 0.0380	0.5729 ± 0.0378	0.5831 ± 0.0339	0.5918 ± 0.0407	0.6116 ± 0.0480	0.6219 ± 0.0485
PRDC Precision	CTGAN	0.9577 ± 0.0095	0.9516 ± 0.0092	0.9458 ± 0.0104	0.9376 ± 0.0140	0.9293 ± 0.0133	0.9200 ± 0.0233	0.9096 ± 0.0288	0.9029 ± 0.0330	0.8963 ± 0.0358	0.8893 ± 0.0323	0.8803 ± 0.0385	0.8508 ± 0.0550	0.8396 ± 0.0656
PRDC Recall	CTGAN	0.9069 ± 0.0156	0.8663 ± 0.0320	0.8503 ± 0.0339	0.8392 ± 0.0386	0.8285 ± 0.0366	0.8239 ± 0.0416	0.8147 ± 0.0621	0.8011 ± 0.0719	0.7965 ± 0.0715	0.7902 ± 0.0601	0.7808 ± 0.0695	0.7576 ± 0.0793	0.7335 ± 0.0805
PRDC Density	CTGAN	0.9467 ± 0.0473	0.9039 ± 0.0392	0.8749 ± 0.0469	0.8401 ± 0.0501	0.8114 ± 0.0578	0.7778 ± 0.0695	0.7484 ± 0.0809	0.7298 ± 0.0824	0.7122 ± 0.0793	0.6914 ± 0.0779	0.6676 ± 0.0845	0.6119 ± 0.0915	0.5899 ± 0.0887
PRDC Coverage	CTGAN	0.9105 ± 0.0349	0.8556 ± 0.0471	0.8246 ± 0.0548	0.7929 ± 0.0598	0.7552 ± 0.0608	0.7293 ± 0.0654	0.7065 ± 0.0864	0.6847 ± 0.0863	0.6710 ± 0.0839	0.6564 ± 0.0780	0.6351 ± 0.0798	0.5658 ± 0.0778	0.5365 ± 0.0657
TV Complement	CTGAN	0.9664 ± 0.0112	0.9490 ± 0.0200	0.9374 ± 0.0166	0.9232 ± 0.0178	0.9113 ± 0.0157	0.8987 ± 0.0168	0.8896 ± 0.0191	0.8780 ± 0.0215	0.8698 ± 0.0213	0.8627 ± 0.0219	0.8570 ± 0.0260	0.8234 ± 0.0314	0.8009 ± 0.0275

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
KS Complement	CTGAN	0.9335 ± 0.0279	0.9048 ± 0.0339	0.8937 ± 0.0322	0.8749 ± 0.0334	0.8646 ± 0.0298	0.8493 ± 0.0361	0.8377 ± 0.0390	0.8291 ± 0.0372	0.8273 ± 0.0390	0.8206 ± 0.0359	0.8119 ± 0.0366	0.7818 ± 0.0413	0.7648 ± 0.0407
Wasserstein Distance	CTGAN	0.1096 ± 0.0043	0.1212 ± 0.0180	0.1309 ± 0.0210	0.1430 ± 0.0225	0.1534 ± 0.0218	0.1663 ± 0.0265	0.1768 ± 0.0301	0.1922 ± 0.0360	0.2040 ± 0.0405	0.2126 ± 0.0420	0.2249 ± 0.0503	0.2777 ± 0.0674	0.3163 ± 0.0678
Maximum Mean Discrepancy	CTGAN	0.0048 ± 0.0047	0.0089 ± 0.0081	0.0108 ± 0.0084	0.0146 ± 0.0088	0.0181 ± 0.0091	0.0237 ± 0.0097	0.0278 ± 0.0120	0.0325 ± 0.0153	0.0360 ± 0.0166	0.0395 ± 0.0166	0.0446 ± 0.0167	0.0666 ± 0.0209	0.0852 ± 0.0275
Jensen-Shannon Divergence	CTGAN	0.0979 ± 0.0196	0.1398 ± 0.0249	0.1651 ± 0.0238	0.1881 ± 0.0228	0.2047 ± 0.0181	0.2187 ± 0.0175	0.2329 ± 0.0194	0.2477 ± 0.0219	0.2567 ± 0.0231	0.2653 ± 0.0206	0.2741 ± 0.0195	0.3075 ± 0.0195	0.3312 ± 0.0187
Detector Average	CTGAN	0.6602 ± 0.0468	0.7216 ± 0.0470	0.7483 ± 0.0380	0.7707 ± 0.0342	0.7858 ± 0.0356	0.8054 ± 0.0278	0.8186 ± 0.0322	0.8312 ± 0.0398	0.8433 ± 0.0338	0.8478 ± 0.0386	0.8597 ± 0.0292	0.9064 ± 0.0301	0.9060 ± 0.0323
GMM Detector	CTGAN	0.5368 ± 0.0166	0.5606 ± 0.0327	0.5879 ± 0.0338	0.5942 ± 0.0366	0.5997 ± 0.0620	0.6261 ± 0.0538	0.6368 ± 0.0588	0.6576 ± 0.0780	0.6811 ± 0.0768	0.6792 ± 0.1084	0.6927 ± 0.0634	0.8034 ± 0.0815	0.7661 ± 0.1008
XGBoost Detector	CTGAN	0.8447 ± 0.0484	0.9298 ± 0.0274	0.9611 ± 0.0143	0.9762 ± 0.0092	0.9848 ± 0.0052	0.9884 ± 0.0047	0.9918 ± 0.0040	0.9941 ± 0.0031	0.9954 ± 0.0026	0.9965 ± 0.0017	0.9973 ± 0.0012	0.9988 ± 0.0007	0.9993 ± 0.0003
MLP Detector	CTGAN	0.6071 ± 0.0808	0.6932 ± 0.0794	0.7247 ± 0.0653	0.7598 ± 0.0484	0.7826 ± 0.0538	0.7993 ± 0.0462	0.8189 ± 0.0479	0.8312 ± 0.0462	0.8443 ± 0.0422	0.8516 ± 0.0513	0.8687 ± 0.0430	0.9041 ± 0.0291	0.9223 ± 0.0272
Linear Detector	CTGAN	0.6522 ± 0.0791	0.7027 ± 0.0779	0.7196 ± 0.0677	0.7526 ± 0.0668	0.7762 ± 0.0633	0.8076 ± 0.0607	0.8268 ± 0.0571	0.8420 ± 0.0576	0.8524 ± 0.0537	0.8637 ± 0.0477	0.8800 ± 0.0434	0.9191 ± 0.0287	0.9362 ± 0.0243
SDMetrics Column Shapes	CTGAN	0.9492 ± 0.0193	0.9260 ± 0.0251	0.9146 ± 0.0228	0.8980 ± 0.0226	0.8870 ± 0.0198	0.8729 ± 0.0225	0.8625 ± 0.0234	0.8525 ± 0.0246	0.8476 ± 0.0247	0.8407 ± 0.0232	0.8335 ± 0.0234	0.8017 ± 0.0250	0.7821 ± 0.0219
SDMetrics Column Pair Trends	CTGAN	0.8693 ± 0.0150	0.8087 ± 0.0234	0.7714 ± 0.0175	0.7482 ± 0.0166	0.7278 ± 0.0166	0.7078 ± 0.0134	0.6948 ± 0.0188	0.6811 ± 0.0210	0.6702 ± 0.0232	0.6593 ± 0.0228	0.6542 ± 0.0220	0.6334 ± 0.0168	0.6298 ± 0.0171
Memorized Fraction	CTGAN	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
New Row Synthesis	CTGAN	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	CTGAN	412.6673 ± 115.0635	391.2847 ± 137.6091	418.9933 ± 124.1791	435.8433 ± 180.2749	400.9393 ± 204.0214	464.5673 ± 104.0775	446.1653 ± 159.6197	363.4893 ± 133.2626	346.7160 ± 143.0422	314.3820 ± 78.2172	400.8407 ± 182.6729	375.5300 ± 182.0659	332.8320 ± 110.0927
Generation Time (s)	CTGAN	0.0607 ± 0.0029	0.0600 ± 0.0000	0.0600 ± 0.0000	0.0607 ± 0.0029	0.0600 ± 0.0000	0.0607 ± 0.0029	0.0607 ± 0.0029	0.0607 ± 0.0029	0.0600 ± 0.0000	0.0600 ± 0.0000	0.0600 ± 0.0000	0.0600 ± 0.0000	0.0613 ± 0.0039
Model Size (MB)	CTGAN	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	CTGAN	0.1318 ± 0.0285	0.1450 ± 0.0304	0.1647 ± 0.0457	0.1861 ± 0.0656	0.2031 ± 0.0719	0.2209 ± 0.0822	0.2376 ± 0.0864	0.2506 ± 0.0944	0.2665 ± 0.0900	0.2764 ± 0.0932	0.2880 ± 0.0968	0.3330 ± 0.0984	0.3557 ± 0.0899
Readmission Collapsed	CTGAN	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
TRTR AUROC	CTGAN	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054
TSTR AUROC	CTGAN	0.6250 ± 0.0113	0.6108 ± 0.0149	0.5925 ± 0.0232	0.5815 ± 0.0292	0.5619 ± 0.0276	0.5567 ± 0.0418	0.5407 ± 0.0386	0.5303 ± 0.0373	0.5314 ± 0.0202	0.5214 ± 0.0325	0.5138 ± 0.0399	0.5084 ± 0.0462	0.4915 ± 0.0376
TSTR Accuracy	CTGAN	0.5836 ± 0.1647	0.5093 ± 0.1713	0.4470 ± 0.2363	0.3474 ± 0.2180	0.2971 ± 0.2203	0.2580 ± 0.2412	0.2540 ± 0.2397	0.2513 ± 0.2354	0.2377 ± 0.2291	0.2344 ± 0.1952	0.2154 ± 0.1604	0.2330 ± 0.2237	0.2067 ± 0.1348
TSTR Precision	CTGAN	0.1803 ± 0.0321	0.1624 ± 0.0177	0.1407 ± 0.0465	0.1402 ± 0.0122	0.1358 ± 0.0107	0.1326 ± 0.0105	0.1207 ± 0.0374	0.1378 ± 0.0333	0.1208 ± 0.0374	0.1291 ± 0.0046	0.1295 ± 0.0079	0.1183 ± 0.0365	0.1255 ± 0.0063
TSTR Recall	CTGAN	0.5727 ± 0.2586	0.6536 ± 0.2290	0.6867 ± 0.2980	0.7843 ± 0.2701	0.8202 ± 0.2623	0.8509 ± 0.3017	0.8436 ± 0.3144	0.8526 ± 0.2986	0.8659 ± 0.2972	0.8649 ± 0.2449	0.8925 ± 0.1905	0.8534 ± 0.3009	0.8798 ± 0.1991
TSTR F1	CTGAN	0.2495 ± 0.0457	0.2502 ± 0.0269	0.2288 ± 0.0715	0.2278 ± 0.0462	0.2266 ± 0.0226	0.2171 ± 0.0497	0.2101 ± 0.0650	0.2150 ± 0.0497	0.2114 ± 0.0649	0.2218 ± 0.0098	0.2245 ± 0.0069	0.2072 ± 0.0642	0.2187 ± 0.0142
TSTR FPR	CTGAN	0.4148 ± 0.2258	0.5117 ± 0.2285	0.5880 ± 0.3130	0.7163 ± 0.2882	0.7792 ± 0.2902	0.8285 ± 0.3200	0.8320 ± 0.3202	0.8364 ± 0.3131	0.8539 ± 0.3056	0.8576 ± 0.2593	0.8834 ± 0.2114	0.8575 ± 0.2999	0.8914 ± 0.1830
TSTR FNR	CTGAN	0.4273 ± 0.2586	0.3464 ± 0.2290	0.3133 ± 0.2980	0.2157 ± 0.2701	0.1798 ± 0.2623	0.1491 ± 0.3017	0.1564 ± 0.3144	0.1474 ± 0.2986	0.1341 ± 0.2972	0.1351 ± 0.2449	0.1075 ± 0.1905	0.1466 ± 0.3009	0.1202 ± 0.1991
TSTR TNR	CTGAN	0.5852 ± 0.2258	0.4883 ± 0.2285	0.4120 ± 0.3130	0.2837 ± 0.2882	0.2208 ± 0.2902	0.1715 ± 0.3200	0.1680 ± 0.3202	0.1636 ± 0.3131	0.1461 ± 0.3056	0.1424 ± 0.2593	0.1166 ± 0.2114	0.1425 ± 0.2999	0.1086 ± 0.1830
alpha-Precision	CTGAN CORE	0.9657 ± 0.0150	0.9665 ± 0.0262	0.9643 ± 0.0257	0.9666 ± 0.0317	0.9668 ± 0.0318	0.9621 ± 0.0410	0.9586 ± 0.0384	0.9519 ± 0.0487	0.9438 ± 0.0601	0.9372 ± 0.0664	0.9277 ± 0.0826	0.8892 ± 0.1022	0.8727 ± 0.1153
beta-Recall	CTGAN CORE	0.4938 ± 0.0163	0.4831 ± 0.0254	0.4718 ± 0.0250	0.4569 ± 0.0240	0.4546 ± 0.0254	0.4447 ± 0.0295	0.4391 ± 0.0295	0.4261 ± 0.0351	0.4164 ± 0.0315	0.4091 ± 0.0322	0.4051 ± 0.0301	0.3762 ± 0.0395	0.3623 ± 0.0435
Authenticity	CTGAN CORE	0.5090 ± 0.0110	0.5161 ± 0.0117	0.5215 ± 0.0140	0.5288 ± 0.0165	0.5305 ± 0.0189	0.5401 ± 0.0240	0.5399 ± 0.0257	0.5511 ± 0.0285	0.5575 ± 0.0262	0.5644 ± 0.0222	0.5715 ± 0.0227	0.5921 ± 0.0325	0.6024 ± 0.0309
PRDC Precision	CTGAN CORE	0.9716 ± 0.0080	0.9651 ± 0.0118	0.9583 ± 0.0139	0.9541 ± 0.0139	0.9479 ± 0.0149	0.9453 ± 0.0147	0.9407 ± 0.0141	0.9379 ± 0.0174	0.9296 ± 0.0186	0.9247 ± 0.0225	0.9195 ± 0.0299	0.9102 ± 0.0361	0.8995 ± 0.0397
PRDC Recall	CTGAN CORE	0.9226 ± 0.0165	0.8768 ± 0.0319	0.8553 ± 0.0365	0.8484 ± 0.0441	0.8475 ± 0.0414	0.8363 ± 0.0334	0.8278 ± 0.0396	0.8137 ± 0.0447	0.8088 ± 0.0429	0.7791 ± 0.0692	0.7748 ± 0.0708	0.7267 ± 0.1060	0.6698 ± 0.1119
PRDC Density	CTGAN CORE	0.9629 ± 0.0454	0.9353 ± 0.0492	0.9013 ± 0.0556	0.8786 ± 0.0514	0.8594 ± 0.0513	0.8465 ± 0.0462	0.8300 ± 0.0425	0.8143 ± 0.0465	0.7912 ± 0.0377	0.7737 ± 0.0414	0.7597 ± 0.0510	0.7250 ± 0.0768	0.6934 ± 0.0804
PRDC Coverage	CTGAN CORE	0.9046 ± 0.0203	0.8530 ± 0.0265	0.8193 ± 0.0283	0.7995 ± 0.0285	0.7848 ± 0.0269	0.7645 ± 0.0290	0.7477 ± 0.0340	0.7237 ± 0.0432	0.7044 ± 0.0391	0.6804 ± 0.0457	0.6695 ± 0.0458	0.6126 ± 0.0781	0.5608 ± 0.0796
TV Complement	CTGAN CORE	0.9704 ± 0.0072	0.9575 ± 0.0138	0.9461 ± 0.0163	0.9418 ± 0.0171	0.9303 ± 0.0181	0.9195 ± 0.0198	0.9154 ± 0.0246	0.9072 ± 0.0274	0.9004 ± 0.0271	0.8947 ± 0.0323	0.8869 ± 0.0297	0.8442 ± 0.0428	0.8120 ± 0.0389
KS Complement	CTGAN CORE	0.9132 ± 0.0224	0.8865 ± 0.0243	0.8740 ± 0.0219	0.8652 ± 0.0249	0.8569 ± 0.0282	0.8483 ± 0.0272	0.8447 ± 0.0264	0.8385 ± 0.0294	0.8309 ± 0.0286	0.8256 ± 0.0317	0.8155 ± 0.0317	0.7961 ± 0.0402	0.7751 ± 0.0426
Wasserstein Distance	CTGAN CORE	0.0201 ± 0.0022	0.0235 ± 0.0027	0.0264 ± 0.0041	0.0292 ± 0.0056	0.0345 ± 0.0072	0.0402 ± 0.0089	0.0444 ± 0.0121	0.0495 ± 0.0137	0.0560 ± 0.0167	0.0619 ± 0.0212	0.0676 ± 0.0244	0.1089 ± 0.0480	0.1511 ± 0.0602
Maximum Mean Discrepancy	CTGAN CORE	0.0065 ± 0.0038	0.0112 ± 0.0049	0.0137 ± 0.0055	0.0163 ± 0.0071	0.0195 ± 0.0100	0.0239 ± 0.0117	0.0263 ± 0.0125	0.0305 ± 0.0135	0.0348 ± 0.0139	0.0393 ± 0.0152	0.0444 ± 0.0170	0.0716 ± 0.0290	0.0949 ± 0.0347

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Jensen-Shannon Divergence	CTGAN CORE	0.1327 ± 0.0177	0.1852 ± 0.0232	0.2162 ± 0.0211	0.2350 ± 0.0262	0.2550 ± 0.0256	0.2742 ± 0.0271	0.2858 ± 0.0325	0.2975 ± 0.0292	0.3058 ± 0.0283	0.3185 ± 0.0315	0.3280 ± 0.0331	0.3707 ± 0.0372	0.4126 ± 0.0349
Detector Average	CTGAN CORE	0.6585 ± 0.0273	0.7003 ± 0.0278	0.7222 ± 0.0229	0.7357 ± 0.0208	0.7505 ± 0.0225	0.7654 ± 0.0205	0.7733 ± 0.0263	0.7840 ± 0.0276	0.8014 ± 0.0253	0.8114 ± 0.0265	0.8143 ± 0.0235	0.8479 ± 0.0430	0.8652 ± 0.0373
GMM Detector	CTGAN CORE	0.5196 ± 0.0175	0.5388 ± 0.0307	0.5475 ± 0.0499	0.5637 ± 0.0416	0.5732 ± 0.0430	0.5920 ± 0.0499	0.5944 ± 0.0456	0.5982 ± 0.0595	0.6420 ± 0.0520	0.6451 ± 0.0589	0.6289 ± 0.0574	0.6695 ± 0.1054	0.6857 ± 0.1054
XGBoost Detector	CTGAN CORE	0.8274 ± 0.0380	0.9118 ± 0.0275	0.9453 ± 0.0205	0.9595 ± 0.0210	0.9719 ± 0.0141	0.9801 ± 0.0106	0.9843 ± 0.0100	0.9882 ± 0.0079	0.9905 ± 0.0065	0.9923 ± 0.0060	0.9933 ± 0.0048	0.9975 ± 0.0016	0.9988 ± 0.0014
MLP Detector	CTGAN CORE	0.6378 ± 0.0449	0.6727 ± 0.0390	0.6998 ± 0.0269	0.7112 ± 0.0336	0.7327 ± 0.0360	0.7511 ± 0.0297	0.7638 ± 0.0299	0.7829 ± 0.0375	0.7940 ± 0.0327	0.8140 ± 0.0281	0.8241 ± 0.0297	0.8634 ± 0.0415	0.8890 ± 0.0373
Linear Detector	CTGAN CORE	0.6492 ± 0.0445	0.6777 ± 0.0456	0.6963 ± 0.0435	0.7082 ± 0.0424	0.7241 ± 0.0518	0.7386 ± 0.0551	0.7507 ± 0.0556	0.7667 ± 0.0607	0.7792 ± 0.0564	0.7943 ± 0.0573	0.8109 ± 0.0572	0.8611 ± 0.0543	0.8874 ± 0.0463
SDMetrics Column Shapes	CTGAN CORE	0.9352 ± 0.0142	0.9138 ± 0.0152	0.9018 ± 0.0147	0.8946 ± 0.0163	0.8851 ± 0.0178	0.8757 ± 0.0168	0.8719 ± 0.0165	0.8649 ± 0.0195	0.8576 ± 0.0192	0.8522 ± 0.0220	0.8430 ± 0.0232	0.8146 ± 0.0322	0.7893 ± 0.0349
SDMetrics Column Pair Trends	CTGAN CORE	0.9591 ± 0.0133	0.9147 ± 0.0202	0.8790 ± 0.0186	0.8486 ± 0.0138	0.8243 ± 0.0125	0.8076 ± 0.0152	0.7897 ± 0.0143	0.7728 ± 0.0119	0.7573 ± 0.0117	0.7432 ± 0.0107	0.7331 ± 0.0139	0.7094 ± 0.0137	0.6977 ± 0.0155
Memorized Fraction	CTGAN CORE	1.3867 ± 0.2642	0.3827 ± 0.1849	0.3667 ± 0.2071	0.3360 ± 0.2098	0.2280 ± 0.1095	0.2087 ± 0.1121	0.1927 ± 0.1218	0.1647 ± 0.1103	0.1553 ± 0.1321	0.1340 ± 0.1125	0.1347 ± 0.1523	0.0660 ± 0.0946	0.0427 ± 0.0651
New Row Synthesis	CTGAN CORE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	CTGAN CORE	273.3100 ± 82.4941	296.6487 ± 80.9920	281.5940 ± 87.0322	233.0427 ± 61.3226	255.3087 ± 77.7327	269.4893 ± 86.5356	284.1413 ± 87.4467	271.8760 ± 94.4076	262.4687 ± 95.5732	295.8847 ± 115.2412	258.7920 ± 98.1325	267.0507 ± 93.2758	303.6707 ± 122.3208
Generation Time (s)	CTGAN CORE	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0567 ± 0.0715	0.0407 ± 0.0029	0.0407 ± 0.0029	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0400 ± 0.0000	0.0407 ± 0.0029	0.0400 ± 0.0000
Model Size (MB)	CTGAN CORE	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	CTGAN CORE	0.1351 ± 0.0159	0.1324 ± 0.0286	0.1348 ± 0.0354	0.1501 ± 0.0342	0.1697 ± 0.0443	0.1813 ± 0.0594	0.1935 ± 0.0544	0.1964 ± 0.0550	0.2176 ± 0.0566	0.2262 ± 0.0634	0.2304 ± 0.0685	0.2962 ± 0.1075	0.3419 ± 0.1030
Readmission Collapsed	CTGAN CORE	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TRTR AUROC	CTGAN CORE	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078
TSTR AUROC	CTGAN CORE	0.5938 ± 0.0133	0.5808 ± 0.0186	0.5759 ± 0.0222	0.5701 ± 0.0166	0.5634 ± 0.0282	0.5385 ± 0.0352	0.5432 ± 0.0283	0.5424 ± 0.0368	0.5285 ± 0.0461	0.5309 ± 0.0427	0.5090 ± 0.0429	0.4995 ± 0.0483	0.4993 ± 0.0508
TSTR Accuracy	CTGAN CORE	0.4568 ± 0.1419	0.5386 ± 0.1656	0.5064 ± 0.2117	0.3821 ± 0.2073	0.2714 ± 0.2291	0.2660 ± 0.2715	0.2474 ± 0.2197	0.2351 ± 0.2315	0.1894 ± 0.1466	0.1816 ± 0.1187	0.1783 ± 0.0963	0.1544 ± 0.0611	0.1397 ± 0.0510
TSTR Precision	CTGAN CORE	0.1529 ± 0.0144	0.1547 ± 0.0135	0.1491 ± 0.0158	0.1422 ± 0.0140	0.1329 ± 0.0086	0.1215 ± 0.0377	0.1312 ± 0.0078	0.1207 ± 0.0372	0.1295 ± 0.0089	0.1303 ± 0.0086	0.1280 ± 0.0019	0.1262 ± 0.0032	0.1278 ± 0.0021

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
TSTR Recall	CTGAN CORE	0.6978 ± 0.1576	0.5689 ± 0.2570	0.5857 ± 0.2754	0.7372 ± 0.2324	0.8408 ± 0.2906	0.8268 ± 0.3612	0.8632 ± 0.2765	0.8699 ± 0.3060	0.9256 ± 0.1683	0.9475 ± 0.1108	0.9372 ± 0.1222	0.9543 ± 0.1062	0.9874 ± 0.0516
TSTR F1	CTGAN CORE	0.2478 ± 0.0084	0.2275 ± 0.0489	0.2273 ± 0.0309	0.2332 ± 0.0159	0.2194 ± 0.0436	0.2042 ± 0.0701	0.2229 ± 0.0207	0.2110 ± 0.0651	0.2253 ± 0.0052	0.2284 ± 0.0098	0.2248 ± 0.0044	0.2227 ± 0.0081	0.2262 ± 0.0015
TSTR FPR	CTGAN CORE	0.5784 ± 0.1854	0.4659 ± 0.2267	0.5051 ± 0.2821	0.6697 ± 0.2710	0.8117 ± 0.3044	0.8158 ± 0.3637	0.8425 ± 0.2917	0.8575 ± 0.3097	0.9180 ± 0.1923	0.9301 ± 0.1517	0.9324 ± 0.1281	0.9623 ± 0.0855	0.9839 ± 0.0659
TSTR FNR	CTGAN CORE	0.3022 ± 0.1576	0.4311 ± 0.2570	0.4143 ± 0.2754	0.2628 ± 0.2324	0.1592 ± 0.2906	0.1732 ± 0.3612	0.1368 ± 0.2765	0.1301 ± 0.3060	0.0744 ± 0.1683	0.0525 ± 0.1108	0.0628 ± 0.1222	0.0457 ± 0.1062	0.0126 ± 0.0516
TSTR TNR	CTGAN CORE	0.4216 ± 0.1854	0.5341 ± 0.2267	0.4949 ± 0.2821	0.3303 ± 0.2710	0.1883 ± 0.3044	0.1842 ± 0.3637	0.1575 ± 0.2917	0.1425 ± 0.3097	0.0820 ± 0.1923	0.0699 ± 0.1517	0.0676 ± 0.1281	0.0377 ± 0.0855	0.0161 ± 0.0659
alpha-Precision	DDPM	0.9571 ± 0.0137	0.7981 ± 0.0431	0.5406 ± 0.0729	0.2567 ± 0.0627	0.1124 ± 0.0408	0.0715 ± 0.0104	0.0693 ± 0.0072	0.0688 ± 0.0071	0.0682 ± 0.0077	0.0680 ± 0.0077	0.0674 ± 0.0078	0.0662 ± 0.0074	0.0657 ± 0.0073
beta-Recall	DDPM	0.4785 ± 0.0093	0.4239 ± 0.0197	0.2971 ± 0.0389	0.1284 ± 0.0398	0.0258 ± 0.0300	0.0017 ± 0.0044	0.0006 ± 0.0006	0.0006 ± 0.0006	0.0006 ± 0.0005	0.0006 ± 0.0006	0.0006 ± 0.0006	0.0006 ± 0.0006	0.0006 ± 0.0006
Authenticity	DDPM	0.5095 ± 0.0079	0.5306 ± 0.0097	0.5867 ± 0.0178	0.6980 ± 0.0424	0.8689 ± 0.0949	0.9916 ± 0.0354	0.9975 ± 0.0093	0.9948 ± 0.0190	0.9910 ± 0.0281	0.9997 ± 0.0008	0.9968 ± 0.0131	0.9999 ± 0.0004	0.9985 ± 0.0055
PRDC Precision	DDPM	0.9325 ± 0.0135	0.8043 ± 0.0459	0.6350 ± 0.0879	0.4935 ± 0.1308	0.4648 ± 0.1424	0.4684 ± 0.1195	0.4740 ± 0.1241	0.4732 ± 0.1269	0.4743 ± 0.1283	0.4744 ± 0.1298	0.4745 ± 0.1303	0.4732 ± 0.1334	0.4712 ± 0.1368
PRDC Recall	DDPM	0.9681 ± 0.0016	0.9656 ± 0.0040	0.9292 ± 0.0109	0.8051 ± 0.0423	0.7566 ± 0.1203	0.4522 ± 0.4064	0.0237 ± 0.0994	0.0000 ± 0.0001	0.0000 ± 0.0000	0.0000 ± 0.0001	0.0000 ± 0.0001	0.0000 ± 0.0000	0.0001 ± 0.0002
PRDC Density	DDPM	0.9377 ± 0.0196	0.7566 ± 0.0494	0.4932 ± 0.0801	0.2514 ± 0.0792	0.1564 ± 0.0683	0.1409 ± 0.0511	0.1421 ± 0.0548	0.1418 ± 0.0550	0.1419 ± 0.0552	0.1421 ± 0.0557	0.1420 ± 0.0557	0.1411 ± 0.0558	0.1403 ± 0.0565
PRDC Coverage	DDPM	0.9586 ± 0.0031	0.9099 ± 0.0142	0.7002 ± 0.0634	0.3089 ± 0.0858	0.0633 ± 0.0615	0.0025 ± 0.0097	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0.0001
TV Complement	DDPM	0.9829 ± 0.0039	0.9239 ± 0.0192	0.8415 ± 0.0428	0.7255 ± 0.0602	0.6773 ± 0.0711	0.6528 ± 0.0622	0.6471 ± 0.0591	0.6446 ± 0.0590	0.6419 ± 0.0587	0.6395 ± 0.0586	0.6370 ± 0.0588	0.6257 ± 0.0595	0.6183 ± 0.0610
KS Complement	DDPM	0.9713 ± 0.0071	0.8749 ± 0.0285	0.7451 ± 0.0547	0.5542 ± 0.0597	0.4612 ± 0.0530	0.4106 ± 0.0355	0.4013 ± 0.0283	0.3999 ± 0.0290	0.3986 ± 0.0300	0.3973 ± 0.0308	0.3959 ± 0.0319	0.3899 ± 0.0353	0.3854 ± 0.0365
Wasserstein Distance	DDPM	0.1723 ± 0.0246	0.4893 ± 0.1105	0.9641 ± 0.2565	1.7370 ± 0.4256	2.1562 ± 0.4692	2.3625 ± 0.3894	2.4091 ± 0.3521	2.4188 ± 0.3565	2.4309 ± 0.3600	2.4382 ± 0.3619	2.4482 ± 0.3630	2.4875 ± 0.3749	2.5135 ± 0.3841
Maximum Mean Discrepancy	DDPM	0.0008 ± 0.0003	0.0192 ± 0.0081	0.0998 ± 0.0308	0.2442 ± 0.0394	0.3371 ± 0.0297	0.3650 ± 0.0055	0.3661 ± 0.0009	0.3661 ± 0.0009	0.3661 ± 0.0009	0.3662 ± 0.0009	0.3662 ± 0.0009	0.3664 ± 0.0010	0.3666 ± 0.0011
Jensen-Shannon Divergence	DDPM	0.0917 ± 0.0120	0.2158 ± 0.0283	0.3463 ± 0.0446	0.5033 ± 0.0458	0.5847 ± 0.0526	0.6447 ± 0.0375	0.6534 ± 0.0262	0.6561 ± 0.0258	0.6577 ± 0.0251	0.6588 ± 0.0253	0.6597 ± 0.0259	0.6685 ± 0.0273	0.6740 ± 0.0281
Detector Average	DDPM	0.5795 ± 0.0045	0.6621 ± 0.0154	0.7565 ± 0.0348	0.8417 ± 0.0425	0.8867 ± 0.0396	0.8783 ± 0.0618	0.8981 ± 0.0068	0.8967 ± 0.0060	0.8888 ± 0.0386	0.8813 ± 0.0512	0.8894 ± 0.0383	0.8705 ± 0.0561	0.8962 ± 0.0076
GMM Detector	DDPM	0.5139 ± 0.0058	0.5383 ± 0.0110	0.5914 ± 0.0275	0.6262 ± 0.0607	0.6356 ± 0.0495	0.6229 ± 0.0523	0.5972 ± 0.0271	0.5914 ± 0.0233	0.5927 ± 0.0304	0.5957 ± 0.0247	0.5944 ± 0.0309	0.5842 ± 0.0232	0.5864 ± 0.0307
XGBoost Detector	DDPM	0.7338 ± 0.0066	0.8614 ± 0.0160	0.9665 ± 0.0120	0.9972 ± 0.0017	0.9999 ± 0.0001	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
MLP Detector	DDPM	0.5252 ± 0.0107	0.6222 ± 0.0246	0.7310 ± 0.1114	0.8625 ± 0.1659	0.9508 ± 0.1393	0.8995 ± 0.2290	1.0000 ± 0.0001	0.9998 ± 0.0003	0.9666 ± 0.1430	0.9332 ± 0.1948	0.9665 ± 0.1429	0.8999 ± 0.2292	0.9998 ± 0.0004
Linear Detector	DDPM	0.5450 ± 0.0062	0.6266 ± 0.0212	0.7371 ± 0.0338	0.8809 ± 0.0360	0.9605 ± 0.0283	0.9908 ± 0.0078	0.9951 ± 0.0023	0.9957 ± 0.0024	0.9959 ± 0.0028	0.9963 ± 0.0028	0.9968 ± 0.0023	0.9977 ± 0.0020	0.9984 ± 0.0018
SDMetrics Column Shapes	DDPM	0.9768 ± 0.0054	0.8983 ± 0.0231	0.7912 ± 0.0478	0.6361 ± 0.0570	0.5645 ± 0.0569	0.5264 ± 0.0392	0.5188 ± 0.0318	0.5169 ± 0.0321	0.5150 ± 0.0321	0.5132 ± 0.0323	0.5112 ± 0.0324	0.5027 ± 0.0335	0.4968 ± 0.0347
SDMetrics Column Pair Trends	DDPM	0.8965 ± 0.0110	0.8125 ± 0.0300	0.7481 ± 0.0444	0.6753 ± 0.0370	0.6350 ± 0.0459	0.6046 ± 0.0362	0.5921 ± 0.0304	0.5916 ± 0.0287	0.5913 ± 0.0290	0.5917 ± 0.0305	0.5916 ± 0.0323	0.5876 ± 0.0405	0.5824 ± 0.0477
Memorized Fraction	DDPM	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
New Row Synthesis	DDPM	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	DDPM	81.1440 ± 9.9727	76.3513 ± 6.6855	77.8000 ± 8.8437	75.8500 ± 6.3236	76.2487 ± 3.8941	75.8473 ± 5.0499	83.5307 ± 9.3147	94.0047 ± 6.1343	98.0480 ± 6.3462	95.5927 ± 10.0303	95.7420 ± 8.1651	98.1073 ± 11.1021	95.4087 ± 10.2953
Generation Time (s)	DDPM	16.2633 ± 0.8909	16.0533 ± 0.5087	16.1133 ± 0.6959	15.9920 ± 0.4506	15.9000 ± 0.1822	16.2493 ± 0.6092	17.9447 ± 3.7853	14.2007 ± 3.5325	15.4840 ± 5.8312	17.2747 ± 6.7877	14.6820 ± 5.0053	13.0500 ± 0.6725	15.6400 ± 5.8322
Model Size (MB)	DDPM	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	DDPM	0.1342 ± 0.0120	0.1731 ± 0.0500	0.2408 ± 0.1252	0.2682 ± 0.1431	0.2631 ± 0.1552	0.2592 ± 0.1603	0.2551 ± 0.1637	0.2458 ± 0.1668	0.2390 ± 0.1707	0.2322 ± 0.1714	0.2278 ± 0.1730	0.2112 ± 0.1783	0.2011 ± 0.1781
Readmission Collapsed	DDPM	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TRTR AUROC	DDPM	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6539 ± 0.0054
TSTR AUROC	DDPM	0.6434 ± 0.0085	0.6312 ± 0.0113	0.6136 ± 0.0172	0.5889 ± 0.0318	0.5737 ± 0.0401	0.5470 ± 0.0419	0.5511 ± 0.0369	0.5414 ± 0.0325	0.5439 ± 0.0295	0.5428 ± 0.0311	0.5388 ± 0.0323	0.5289 ± 0.0405	0.5321 ± 0.0366
TSTR Accuracy	DDPM	0.6437 ± 0.0460	0.6831 ± 0.0901	0.7187 ± 0.1004	0.7242 ± 0.1080	0.7723 ± 0.0876	0.6916 ± 0.2125	0.7647 ± 0.1433	0.7481 ± 0.1677	0.7307 ± 0.2170	0.7565 ± 0.1560	0.7618 ± 0.1643	0.7353 ± 0.2049	0.7217 ± 0.2698
TSTR Precision	DDPM	0.1875 ± 0.0073	0.1904 ± 0.0163	0.1870 ± 0.0203	0.1726 ± 0.0303	0.1761 ± 0.0815	0.1568 ± 0.0395	0.1507 ± 0.0541	0.2169 ± 0.2419	0.1514 ± 0.0530	0.1529 ± 0.0532	0.1746 ± 0.1468	0.1131 ± 0.0674	0.1100 ± 0.0675
TSTR Recall	DDPM	0.5362 ± 0.0779	0.4397 ± 0.1636	0.3473 ± 0.2190	0.3073 ± 0.2185	0.2063 ± 0.1770	0.2993 ± 0.2955	0.1973 ± 0.2360	0.2090 ± 0.2498	0.2354 ± 0.3042	0.1954 ± 0.2447	0.1820 ± 0.2560	0.2052 ± 0.3055	0.2107 ± 0.3707
TSTR F1	DDPM	0.2767 ± 0.0080	0.2558 ± 0.0294	0.2100 ± 0.0947	0.1825 ± 0.1158	0.1522 ± 0.1064	0.1642 ± 0.0900	0.1344 ± 0.0977	0.1325 ± 0.0937	0.1367 ± 0.0966	0.1210 ± 0.0972	0.1082 ± 0.1029	0.1044 ± 0.0956	0.0948 ± 0.0901
TSTR FPR	DDPM	0.3406 ± 0.0638	0.2813 ± 0.1267	0.2271 ± 0.1469	0.2150 ± 0.1553	0.1451 ± 0.1259	0.2512 ± 0.2860	0.1525 ± 0.1980	0.1732 ± 0.2281	0.1970 ± 0.2925	0.1617 ± 0.2139	0.1536 ± 0.2252	0.1874 ± 0.2792	0.2037 ± 0.3632
TSTR FNR	DDPM	0.4638 ± 0.0779	0.5603 ± 0.1636	0.6527 ± 0.2190	0.6927 ± 0.2185	0.7937 ± 0.1770	0.7007 ± 0.2955	0.8027 ± 0.2360	0.7910 ± 0.2498	0.7646 ± 0.3042	0.8046 ± 0.2447	0.8180 ± 0.2560	0.7948 ± 0.3055	0.7893 ± 0.3707
TSTR TNR	DDPM	0.6594 ± 0.0638	0.7187 ± 0.1267	0.7729 ± 0.1469	0.7850 ± 0.1553	0.8549 ± 0.1259	0.7488 ± 0.2860	0.8475 ± 0.1980	0.8268 ± 0.2281	0.8030 ± 0.2925	0.8383 ± 0.2139	0.8464 ± 0.2252	0.8126 ± 0.2792	0.7963 ± 0.3632

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
alpha-Precision	DDPM CORE	0.9853 ± 0.0055	0.9686 ± 0.0145	0.9318 ± 0.0250	0.8653 ± 0.0371	0.7549 ± 0.0553	0.6404 ± 0.0683	0.5284 ± 0.0716	0.4526 ± 0.0650	0.3891 ± 0.0596	0.3384 ± 0.0589	0.2951 ± 0.0570	0.2133 ± 0.0592	0.2114 ± 0.0609
beta-Recall	DDPM CORE	0.4999 ± 0.0068	0.4919 ± 0.0072	0.4747 ± 0.0116	0.4412 ± 0.0201	0.3707 ± 0.0353	0.2842 ± 0.0405	0.1982 ± 0.0394	0.1365 ± 0.0309	0.0921 ± 0.0232	0.0612 ± 0.0169	0.0407 ± 0.0145	0.0052 ± 0.0061	0.0043 ± 0.0050
Authenticity	DDPM CORE	0.5017 ± 0.0064	0.5044 ± 0.0063	0.5123 ± 0.0079	0.5303 ± 0.0099	0.5578 ± 0.0156	0.6032 ± 0.0218	0.6648 ± 0.0265	0.7240 ± 0.0282	0.7804 ± 0.0280	0.8321 ± 0.0231	0.8661 ± 0.0268	0.9888 ± 0.0211	0.9922 ± 0.0144
PRDC Precision	DDPM CORE	0.9708 ± 0.0031	0.9489 ± 0.0073	0.9049 ± 0.0186	0.8244 ± 0.0348	0.7006 ± 0.0590	0.5735 ± 0.0766	0.4496 ± 0.0876	0.3602 ± 0.0825	0.2859 ± 0.0745	0.2304 ± 0.0766	0.1764 ± 0.0813	0.0372 ± 0.0644	0.0310 ± 0.0578
PRDC Recall	DDPM CORE	0.9813 ± 0.0020	0.9791 ± 0.0015	0.9754 ± 0.0027	0.9636 ± 0.0039	0.9343 ± 0.0098	0.8704 ± 0.0297	0.7471 ± 0.0541	0.5855 ± 0.0753	0.4383 ± 0.0715	0.3478 ± 0.0889	0.3694 ± 0.1485	0.0506 ± 0.1185	0.0273 ± 0.1016
PRDC Density	DDPM CORE	0.9825 ± 0.0056	0.9532 ± 0.0111	0.8951 ± 0.0256	0.7990 ± 0.0418	0.6527 ± 0.0612	0.4992 ± 0.0749	0.3614 ± 0.0765	0.2635 ± 0.0634	0.1874 ± 0.0483	0.1329 ± 0.0441	0.0864 ± 0.0399	0.0098 ± 0.0175	0.0077 ± 0.0147
PRDC Coverage	DDPM CORE	0.9642 ± 0.0019	0.9555 ± 0.0027	0.9376 ± 0.0071	0.8969 ± 0.0160	0.7994 ± 0.0384	0.6233 ± 0.0723	0.4134 ± 0.0817	0.2387 ± 0.0662	0.1225 ± 0.0419	0.0582 ± 0.0254	0.0276 ± 0.0127	0.0009 ± 0.0016	0.0007 ± 0.0012
TV Complement	DDPM CORE	0.9936 ± 0.0011	0.9888 ± 0.0019	0.9798 ± 0.0047	0.9616 ± 0.0128	0.9348 ± 0.0251	0.9105 ± 0.0345	0.8897 ± 0.0432	0.8859 ± 0.0480	0.8870 ± 0.0471	0.8891 ± 0.0518	0.8876 ± 0.0606	0.8878 ± 0.0552	0.8917 ± 0.0530
KS Complement	DDPM CORE	0.9899 ± 0.0015	0.9794 ± 0.0033	0.9609 ± 0.0094	0.9228 ± 0.0248	0.8575 ± 0.0422	0.7926 ± 0.0575	0.7242 ± 0.0679	0.6851 ± 0.0636	0.6530 ± 0.0558	0.6268 ± 0.0586	0.5987 ± 0.0672	0.4876 ± 0.0776	0.4803 ± 0.0782
Wasserstein Distance	DDPM CORE	0.0209 ± 0.0016	0.0349 ± 0.0047	0.0608 ± 0.0145	0.1201 ± 0.0327	0.2283 ± 0.0681	0.3349 ± 0.0879	0.4414 ± 0.1050	0.5039 ± 0.0960	0.5542 ± 0.0861	0.6020 ± 0.0991	0.6522 ± 0.1277	0.8867 ± 0.1918	0.9082 ± 0.2044
Maximum Mean Discrepancy	DDPM CORE	0.0001 ± 0.0001	0.0003 ± 0.0001	0.0020 ± 0.0011	0.0094 ± 0.0049	0.0330 ± 0.0148	0.0737 ± 0.0276	0.1301 ± 0.0377	0.1779 ± 0.0359	0.2217 ± 0.0321	0.2609 ± 0.0327	0.2965 ± 0.0347	0.3711 ± 0.0191	0.3728 ± 0.0174
Jensen-Shannon Divergence	DDPM CORE	0.0584 ± 0.0033	0.0912 ± 0.0076	0.1329 ± 0.0159	0.1930 ± 0.0263	0.2706 ± 0.0375	0.3413 ± 0.0422	0.4082 ± 0.0459	0.4560 ± 0.0430	0.4971 ± 0.0381	0.5315 ± 0.0390	0.5599 ± 0.0413	0.6325 ± 0.0359	0.6382 ± 0.0293
Detector Average	DDPM CORE	0.5235 ± 0.0033	0.5482 ± 0.0051	0.5847 ± 0.0116	0.6294 ± 0.0153	0.6791 ± 0.0244	0.7139 ± 0.0411	0.7512 ± 0.0561	0.7763 ± 0.0524	0.7856 ± 0.0634	0.7999 ± 0.0722	0.8039 ± 0.0761	0.7922 ± 0.0764	0.7673 ± 0.0482
GMM Detector	DDPM CORE	0.5121 ± 0.0099	0.5125 ± 0.0100	0.5164 ± 0.0164	0.5255 ± 0.0145	0.5474 ± 0.0230	0.5584 ± 0.0275	0.5965 ± 0.0537	0.6116 ± 0.0383	0.6156 ± 0.0405	0.6297 ± 0.0489	0.6353 ± 0.0549	0.6323 ± 0.0761	0.6262 ± 0.0807
XGBoost Detector	DDPM CORE	0.5617 ± 0.0060	0.6323 ± 0.0086	0.7149 ± 0.0147	0.8081 ± 0.0212	0.8967 ± 0.0234	0.9581 ± 0.0177	0.9885 ± 0.0076	0.9977 ± 0.0019	0.9997 ± 0.0004	1.0000 ± 0.0001	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
MLP Detector	DDPM CORE	0.5029 ± 0.0032	0.5183 ± 0.0065	0.5631 ± 0.0144	0.6190 ± 0.0157	0.6640 ± 0.0567	0.6837 ± 0.1297	0.7098 ± 0.1717	0.7499 ± 0.2035	0.7506 ± 0.2355	0.7639 ± 0.2475	0.7480 ± 0.2662	0.6324 ± 0.2518	0.5329 ± 0.1414
Linear Detector	DDPM CORE	0.5171 ± 0.0089	0.5298 ± 0.0089	0.5443 ± 0.0129	0.5649 ± 0.0216	0.6081 ± 0.0347	0.6552 ± 0.0419	0.7101 ± 0.0443	0.7461 ± 0.0411	0.7766 ± 0.0427	0.8062 ± 0.0465	0.8323 ± 0.0496	0.9040 ± 0.0587	0.9102 ± 0.0611
SDMetrics Column Shapes	DDPM CORE	0.9913 ± 0.0011	0.9830 ± 0.0023	0.9682 ± 0.0072	0.9378 ± 0.0197	0.8872 ± 0.0342	0.8380 ± 0.0459	0.7879 ± 0.0544	0.7623 ± 0.0529	0.7430 ± 0.0471	0.7277 ± 0.0502	0.7098 ± 0.0586	0.6415 ± 0.0594	0.6386 ± 0.0582
SDMetrics Column Pair Trends	DDPM CORE	0.9960 ± 0.0027	0.9867 ± 0.0072	0.9721 ± 0.0082	0.9539 ± 0.0185	0.9385 ± 0.0202	0.9366 ± 0.0261	0.9334 ± 0.0356	0.9321 ± 0.0367	0.9279 ± 0.0412	0.9241 ± 0.0464	0.9159 ± 0.0490	0.8426 ± 0.0662	0.8326 ± 0.0731

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Memorized Fraction	DDPM CORE	1.4727 ± 0.1260	2.3113 ± 0.1959	2.2247 ± 0.1891	1.9500 ± 0.2170	1.5053 ± 0.3545	1.1187 ± 0.2591	1.1520 ± 0.2835	1.3340 ± 0.3877	1.4687 ± 0.4106	1.2300 ± 0.4279	0.7453 ± 0.4521	0.0087 ± 0.0236	0.0053 ± 0.0138
New Row Synthesis	DDPM CORE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	DDPM CORE	72.5067 ± 8.1038	71.3460 ± 9.4411	68.8733 ± 5.0218	69.7947 ± 7.4224	67.0480 ± 2.3832	67.1060 ± 2.2605	69.7027 ± 6.3036	67.9593 ± 3.7225	67.5353 ± 2.2787	67.8207 ± 2.2608	68.1367 ± 2.9361	74.0687 ± 4.6755	79.8080 ± 4.1513
Generation Time (s)	DDPM CORE	12.6720 ± 0.5115	12.6127 ± 0.5009	12.4607 ± 0.2657	12.5367 ± 0.4197	12.0127 ± 1.7794	12.0273 ± 1.7622	10.8527 ± 3.1716	10.3073 ± 3.3234	11.9927 ± 1.7432	12.4220 ± 0.1183	11.6093 ± 2.4387	12.0953 ± 3.4135	10.3793 ± 3.0471
Model Size (MB)	DDPM CORE	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	DDPM CORE	0.1246 ± 0.0039	0.1287 ± 0.0094	0.1369 ± 0.0182	0.1555 ± 0.0416	0.1805 ± 0.0739	0.2048 ± 0.1004	0.2240 ± 0.1210	0.2202 ± 0.1275	0.2069 ± 0.1240	0.1996 ± 0.1248	0.1940 ± 0.1302	0.1534 ± 0.1226	0.1341 ± 0.1247
Readmission Collapsed	DDPM CORE	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
TRTR AUROC	DDPM CORE	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6162 ± 0.0078
TSTR AUROC	DDPM CORE	0.6047 ± 0.0092	0.5905 ± 0.0118	0.5726 ± 0.0174	0.5533 ± 0.0290	0.5375 ± 0.0283	0.5214 ± 0.0329	0.5222 ± 0.0343	0.5133 ± 0.0325	0.5060 ± 0.0400	0.5187 ± 0.0379	0.5141 ± 0.0356	0.5158 ± 0.0548	0.5134 ± 0.0321
TSTR Accuracy	DDPM CORE	0.5537 ± 0.0845	0.5377 ± 0.0875	0.5226 ± 0.1640	0.5703 ± 0.1426	0.5457 ± 0.2676	0.6083 ± 0.2033	0.6185 ± 0.2163	0.6249 ± 0.2306	0.6348 ± 0.2210	0.6591 ± 0.2012	0.7158 ± 0.1317	0.6974 ± 0.2423	0.7417 ± 0.2664
TSTR Precision	DDPM CORE	0.1642 ± 0.0093	0.1572 ± 0.0084	0.1520 ± 0.0139	0.1498 ± 0.0196	0.1435 ± 0.0177	0.1401 ± 0.0165	0.1399 ± 0.0244	0.1262 ± 0.0427	0.1362 ± 0.0171	0.1232 ± 0.0443	0.1427 ± 0.0215	0.1248 ± 0.0442	0.0754 ± 0.0837
TSTR Recall	DDPM CORE	0.6037 ± 0.1170	0.5961 ± 0.1177	0.5803 ± 0.2007	0.4883 ± 0.1937	0.4893 ± 0.3615	0.3895 ± 0.2705	0.3748 ± 0.2882	0.3499 ± 0.3038	0.3314 ± 0.2915	0.3045 ± 0.2708	0.2253 ± 0.1692	0.2457 ± 0.3459	0.1850 ± 0.3660
TSTR F1	DDPM CORE	0.2559 ± 0.0069	0.2470 ± 0.0064	0.2359 ± 0.0145	0.2208 ± 0.0244	0.1898 ± 0.0747	0.1841 ± 0.0605	0.1805 ± 0.0612	0.1595 ± 0.0834	0.1616 ± 0.0683	0.1594 ± 0.0780	0.1502 ± 0.0540	0.1209 ± 0.0881	0.0732 ± 0.1043
TSTR FPR	DDPM CORE	0.4536 ± 0.1137	0.4708 ± 0.1172	0.4858 ± 0.2166	0.4177 ± 0.1910	0.4461 ± 0.3589	0.3598 ± 0.2720	0.3459 ± 0.2894	0.3350 ± 0.3080	0.3209 ± 0.2951	0.2891 ± 0.2695	0.2127 ± 0.1752	0.2367 ± 0.3273	0.1771 ± 0.3582
TSTR FNR	DDPM CORE	0.3963 ± 0.1170	0.4039 ± 0.1177	0.4197 ± 0.2007	0.5117 ± 0.1937	0.5107 ± 0.3615	0.6105 ± 0.2705	0.6252 ± 0.2882	0.6501 ± 0.3038	0.6686 ± 0.2915	0.6955 ± 0.2708	0.7747 ± 0.1692	0.7543 ± 0.3459	0.8150 ± 0.3660
TSTR TNR	DDPM CORE	0.5464 ± 0.1137	0.5292 ± 0.1172	0.5142 ± 0.2166	0.5823 ± 0.1910	0.5539 ± 0.3589	0.6402 ± 0.2720	0.6541 ± 0.2894	0.6650 ± 0.3080	0.6791 ± 0.2951	0.7109 ± 0.2695	0.7873 ± 0.1752	0.7633 ± 0.3273	0.8229 ± 0.3582
alpha-Precision	RTVAE	0.8072 ± 0.1313	0.7600 ± 0.1359	0.7291 ± 0.1477	0.6867 ± 0.1420	0.6514 ± 0.1460	0.6140 ± 0.1474	0.5907 ± 0.1580	0.5644 ± 0.1622	0.5425 ± 0.1687	0.5267 ± 0.1763	0.5128 ± 0.1760	0.4805 ± 0.1877	0.4740 ± 0.1904
beta-Recall	RTVAE	0.4363 ± 0.1701	0.4094 ± 0.1617	0.3775 ± 0.1498	0.3403 ± 0.1340	0.3059 ± 0.1231	0.2737 ± 0.1148	0.2425 ± 0.1021	0.2153 ± 0.0945	0.1932 ± 0.0879	0.1752 ± 0.0817	0.1585 ± 0.0768	0.1030 ± 0.0580	0.0780 ± 0.0446
Authenticity	RTVAE	0.5642 ± 0.1241	0.5756 ± 0.1169	0.5961 ± 0.1115	0.6231 ± 0.1115	0.6452 ± 0.1049	0.6742 ± 0.1069	0.6928 ± 0.1013	0.7153 ± 0.0964	0.7392 ± 0.0997	0.7570 ± 0.0947	0.7698 ± 0.0887	0.8225 ± 0.0778	0.8465 ± 0.0682
PRDC Precision	RTVAE	0.9519 ± 0.0828	0.9636 ± 0.0639	0.9702 ± 0.0534	0.9736 ± 0.0515	0.9756 ± 0.0502	0.9797 ± 0.0459	0.9819 ± 0.0418	0.9831 ± 0.0433	0.9851 ± 0.0385	0.9868 ± 0.0355	0.9895 ± 0.0271	0.9952 ± 0.0126	0.9953 ± 0.0128
PRDC Recall	RTVAE	0.3570 ± 0.1430	0.2684 ± 0.1323	0.2321 ± 0.1232	0.1730 ± 0.0799	0.1150 ± 0.0705	0.0837 ± 0.0642	0.0630 ± 0.0461	0.0309 ± 0.0302	0.0226 ± 0.0306	0.0129 ± 0.0179	0.0090 ± 0.0162	0.0083 ± 0.0217	0.0027 ± 0.0071

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
PRDC Density	RTVAE	1.0869 ± 0.3519	1.1318 ± 0.3349	1.1710 ± 0.3319	1.2128 ± 0.3425	1.2404 ± 0.3497	1.2642 ± 0.3543	1.2887 ± 0.3669	1.3035 ± 0.3828	1.3090 ± 0.3866	1.3211 ± 0.3979	1.3313 ± 0.3972	1.3529 ± 0.4346	1.3550 ± 0.4697
PRDC Coverage	RTVAE	0.5265 ± 0.2214	0.4448 ± 0.2056	0.3718 ± 0.1733	0.3021 ± 0.1413	0.2435 ± 0.1165	0.1970 ± 0.0981	0.1574 ± 0.0797	0.1286 ± 0.0671	0.1082 ± 0.0574	0.0930 ± 0.0508	0.0809 ± 0.0455	0.0438 ± 0.0338	0.0280 ± 0.0221
TV Complement	RTVAE	0.9083 ± 0.0273	0.8586 ± 0.0322	0.8198 ± 0.0409	0.7919 ± 0.0415	0.7650 ± 0.0456	0.7487 ± 0.0473	0.7380 ± 0.0498	0.7216 ± 0.0442	0.7085 ± 0.0426	0.6966 ± 0.0394	0.6867 ± 0.0400	0.6568 ± 0.0439	0.6460 ± 0.0439
KS Complement	RTVAE	0.7775 ± 0.0688	0.7204 ± 0.0811	0.6814 ± 0.0764	0.6493 ± 0.0712	0.6198 ± 0.0668	0.5966 ± 0.0627	0.5741 ± 0.0592	0.5553 ± 0.0596	0.5401 ± 0.0585	0.5263 ± 0.0571	0.5150 ± 0.0572	0.4821 ± 0.0523	0.4704 ± 0.0492
Wasserstein Distance	RTVAE	0.1288 ± 0.0410	0.1637 ± 0.0401	0.2145 ± 0.0443	0.2542 ± 0.0552	0.2926 ± 0.0669	0.3229 ± 0.0734	0.3520 ± 0.0769	0.3885 ± 0.0665	0.4227 ± 0.0616	0.4499 ± 0.0508	0.4710 ± 0.0507	0.5310 ± 0.0636	0.5541 ± 0.0846
Maximum Mean Discrepancy	RTVAE	0.0434 ± 0.0427	0.0681 ± 0.0434	0.0919 ± 0.0415	0.1110 ± 0.0466	0.1338 ± 0.0526	0.1506 ± 0.0575	0.1628 ± 0.0641	0.1795 ± 0.0645	0.1954 ± 0.0654	0.2077 ± 0.0567	0.2191 ± 0.0535	0.2563 ± 0.0519	0.2654 ± 0.0605
Jensen-Shannon Divergence	RTVAE	0.3984 ± 0.0592	0.4806 ± 0.0210	0.5103 ± 0.0198	0.5312 ± 0.0210	0.5515 ± 0.0217	0.5663 ± 0.0222	0.5783 ± 0.0234	0.5901 ± 0.0196	0.6009 ± 0.0200	0.6087 ± 0.0152	0.6161 ± 0.0158	0.6397 ± 0.0158	0.6465 ± 0.0156
Detector Average	RTVAE	0.7959 ± 0.0682	0.8475 ± 0.0681	0.8720 ± 0.0452	0.9134 ± 0.0615	0.9171 ± 0.0703	0.9421 ± 0.0497	0.9438 ± 0.0487	0.9567 ± 0.0303	0.9442 ± 0.0509	0.9720 ± 0.0286	0.9697 ± 0.0322	0.9676 ± 0.0415	0.9669 ± 0.0385
GMM Detector	RTVAE	0.5618 ± 0.0906	0.6274 ± 0.1821	0.6378 ± 0.1673	0.7590 ± 0.2259	0.7378 ± 0.2488	0.8206 ± 0.1837	0.8159 ± 0.1979	0.8569 ± 0.1250	0.7976 ± 0.2032	0.9030 ± 0.1132	0.8911 ± 0.1263	0.8750 ± 0.1663	0.8706 ± 0.1535
XGBoost Detector	RTVAE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
MLP Detector	RTVAE	0.8221 ± 0.1098	0.8855 ± 0.0712	0.9242 ± 0.0487	0.9464 ± 0.0329	0.9639 ± 0.0226	0.9722 ± 0.0192	0.9787 ± 0.0149	0.9824 ± 0.0129	0.9876 ± 0.0097	0.9912 ± 0.0073	0.9921 ± 0.0070	0.9969 ± 0.0033	0.9979 ± 0.0021
Linear Detector	RTVAE	0.7997 ± 0.1398	0.8773 ± 0.0855	0.9259 ± 0.0539	0.9484 ± 0.0413	0.9666 ± 0.0294	0.9757 ± 0.0220	0.9808 ± 0.0183	0.9875 ± 0.0115	0.9915 ± 0.0079	0.9937 ± 0.0065	0.9955 ± 0.0048	0.9984 ± 0.0020	0.9990 ± 0.0015
SDMetrics Column Shapes	RTVAE	0.8401 ± 0.0450	0.7865 ± 0.0543	0.7476 ± 0.0470	0.7175 ± 0.0469	0.6893 ± 0.0459	0.6694 ± 0.0449	0.6525 ± 0.0459	0.6348 ± 0.0430	0.6206 ± 0.0411	0.6077 ± 0.0374	0.5971 ± 0.0349	0.5657 ± 0.0351	0.5544 ± 0.0340
SDMetrics Column Pair Trends	RTVAE	0.7815 ± 0.0598	0.6984 ± 0.0581	0.6412 ± 0.0574	0.6012 ± 0.0573	0.5669 ± 0.0571	0.5424 ± 0.0737	0.5307 ± 0.0787	0.5103 ± 0.0715	0.4847 ± 0.0782	0.4721 ± 0.0738	0.4464 ± 0.0606	0.3964 ± 0.1007	0.3713 ± 0.1137
Memorized Fraction	RTVAE	0.0013 ± 0.0039	0.0000 ± 0.0000	0.0007 ± 0.0029	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
New Row Synthesis	RTVAE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	RTVAE	145.2693 ± 29.5286	147.5513 ± 24.3507	183.5093 ± 26.1348	181.2713 ± 20.1379	192.5527 ± 22.5302	198.8520 ± 29.4543	207.2533 ± 36.1875	203.9153 ± 23.5121	202.4727 ± 27.1863	218.7873 ± 20.1784	231.2200 ± 28.1936	250.1733 ± 60.7908	278.2800 ± 46.5755
Generation Time (s)	RTVAE	1.0253 ± 1.0890	0.6960 ± 1.0058	0.2807 ± 0.4857	0.1767 ± 0.0589	0.4733 ± 0.6864	0.3507 ± 0.5080	0.5673 ± 0.6984	0.3653 ± 0.5056	0.4960 ± 0.7057	0.6067 ± 0.8559	0.4907 ± 0.9049	0.6413 ± 0.8481	0.5553 ± 1.0966
Model Size (MB)	RTVAE	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Minority class share (readm)	RTVAE	0.0767 ± 0.0179	0.0308 ± 0.0179	0.0048 ± 0.0062	0.0002 ± 0.0010	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Readmission Collapsed	RTVAE	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0667 ± 0.2860	0.8667 ± 0.3897	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
TRTR AUROC	RTVAE	0.6539 ± 0.0054	0.6539 ± 0.0054	0.6540 ± 0.0058	0.6560 ± 0.1127	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR AUROC	RTVAE	0.6071 ± 0.0244	0.5911 ± 0.0250	0.5356 ± 0.0441	0.5333 ± 0.8451	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR Accuracy	RTVAE	0.6958 ± 0.1555	0.7400 ± 0.2327	0.7072 ± 0.3548	0.6452 ± 5.7813	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR Precision	RTVAE	0.1842 ± 0.0321	0.1128 ± 0.0915	0.0397 ± 0.0766	0.0750 ± 1.9059	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR Recall	RTVAE	0.3515 ± 0.2540	0.2269 ± 0.3715	0.2270 ± 0.4705	0.3830 ± 9.7318	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR F1	RTVAE	0.2051 ± 0.0707	0.0987 ± 0.1215	0.0634 ± 0.1201	0.1254 ± 3.1876	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR FPR	RTVAE	0.2540 ± 0.2146	0.1852 ± 0.3204	0.2227 ± 0.4751	0.3165 ± 8.0442	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR FNR	RTVAE	0.6485 ± 0.2540	0.7731 ± 0.3715	0.7730 ± 0.4705	0.6170 ± 9.7318	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR TNR	RTVAE	0.7460 ± 0.2146	0.8148 ± 0.3204	0.7773 ± 0.4751	0.6835 ± 8.0442	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
alpha-Precision	RTVAE CORE	0.7568 ± 0.2787	0.7387 ± 0.2900	0.7215 ± 0.2871	0.7131 ± 0.2873	0.7133 ± 0.2862	0.7036 ± 0.2842	0.6992 ± 0.2821	0.6956 ± 0.2844	0.6888 ± 0.2835	0.6885 ± 0.2837	0.6848 ± 0.2881	0.6816 ± 0.2866	0.6933 ± 0.2928
beta-Recall	RTVAE CORE	0.4063 ± 0.2417	0.3917 ± 0.2354	0.3789 ± 0.2288	0.3664 ± 0.2219	0.3537 ± 0.2141	0.3438 ± 0.2072	0.3336 ± 0.2028	0.3249 ± 0.1984	0.3162 ± 0.1935	0.3090 ± 0.1894	0.3072 ± 0.1870	0.2805 ± 0.1715	0.2686 ± 0.1645
Authenticity	RTVAE CORE	0.6060 ± 0.2224	0.6187 ± 0.2210	0.6282 ± 0.2158	0.6318 ± 0.2035	0.6424 ± 0.2070	0.6535 ± 0.2015	0.6610 ± 0.2015	0.6649 ± 0.2007	0.6740 ± 0.1967	0.6787 ± 0.1937	0.6826 ± 0.1912	0.6995 ± 0.1766	0.7134 ± 0.1751
PRDC Precision	RTVAE CORE	0.8588 ± 0.2708	0.8640 ± 0.2542	0.8652 ± 0.2563	0.8720 ± 0.2458	0.8736 ± 0.2429	0.8746 ± 0.2448	0.8730 ± 0.2482	0.8743 ± 0.2474	0.8793 ± 0.2393	0.8831 ± 0.2306	0.8837 ± 0.2337	0.8851 ± 0.2311	0.8862 ± 0.2361
PRDC Recall	RTVAE CORE	0.5149 ± 0.1806	0.4145 ± 0.2622	0.3868 ± 0.2372	0.3380 ± 0.2291	0.2974 ± 0.1943	0.2554 ± 0.1794	0.2191 ± 0.1483	0.1927 ± 0.1359	0.1757 ± 0.1256	0.1612 ± 0.1112	0.1411 ± 0.1065	0.0931 ± 0.0730	0.0738 ± 0.0614
PRDC Density	RTVAE CORE	0.8982 ± 0.4816	0.8916 ± 0.4732	0.8955 ± 0.4716	0.9023 ± 0.4660	0.9068 ± 0.4674	0.9126 ± 0.4700	0.9066 ± 0.4709	0.9112 ± 0.4685	0.9159 ± 0.4714	0.9213 ± 0.4694	0.9248 ± 0.4699	0.9347 ± 0.4759	0.9319 ± 0.4743
PRDC Coverage	RTVAE CORE	0.5588 ± 0.3556	0.5102 ± 0.3322	0.4636 ± 0.3022	0.4182 ± 0.2742	0.3826 ± 0.2518	0.3510 ± 0.2308	0.3236 ± 0.2136	0.3016 ± 0.2005	0.2822 ± 0.1884	0.2640 ± 0.1769	0.2512 ± 0.1696	0.1982 ± 0.1379	0.1786 ± 0.1245
TV Complement	RTVAE CORE	0.9246 ± 0.0158	0.9004 ± 0.0205	0.8797 ± 0.0227	0.8661 ± 0.0268	0.8547 ± 0.0338	0.8501 ± 0.0385	0.8425 ± 0.0366	0.8332 ± 0.0371	0.8259 ± 0.0390	0.8198 ± 0.0382	0.8213 ± 0.0439	0.8225 ± 0.0522	0.8256 ± 0.0498
KS Complement	RTVAE CORE	0.7006 ± 0.2561	0.6704 ± 0.2498	0.6480 ± 0.2413	0.6302 ± 0.2335	0.6174 ± 0.2284	0.6065 ± 0.2232	0.5966 ± 0.2198	0.5878 ± 0.2161	0.5787 ± 0.2126	0.5718 ± 0.2091	0.5660 ± 0.2062	0.5421 ± 0.1945	0.5331 ± 0.1921
Wasserstein Distance	RTVAE CORE	0.0531 ± 0.0507	0.0686 ± 0.0519	0.0893 ± 0.0590	0.1166 ± 0.0591	0.1306 ± 0.0513	0.1376 ± 0.0510	0.1427 ± 0.0564	0.1464 ± 0.0560	0.1503 ± 0.0570	0.1549 ± 0.0576	0.1551 ± 0.0601	0.1590 ± 0.0632	0.1630 ± 0.0609

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
Maximum Mean Discrepancy	RTVAE CORE	0.1080 ± 0.1765	0.1287 ± 0.1827	0.1475 ± 0.1908	0.1618 ± 0.1930	0.1707 ± 0.1941	0.1795 ± 0.1986	0.1878 ± 0.2035	0.1943 ± 0.2030	0.2015 ± 0.2041	0.2089 ± 0.2020	0.2101 ± 0.2026	0.2202 ± 0.1976	0.2306 ± 0.1954
Jensen-Shannon Divergence	RTVAE CORE	0.5226 ± 0.0487	0.5612 ± 0.0403	0.5760 ± 0.0402	0.5876 ± 0.0373	0.5948 ± 0.0346	0.5985 ± 0.0335	0.6036 ± 0.0345	0.6096 ± 0.0326	0.6133 ± 0.0328	0.6175 ± 0.0311	0.6181 ± 0.0323	0.6261 ± 0.0296	0.6296 ± 0.0266
Detector Average	RTVAE CORE	0.7681 ± 0.1179	0.7929 ± 0.0973	0.8327 ± 0.0925	0.8666 ± 0.0640	0.8765 ± 0.0696	0.9012 ± 0.0530	0.9028 ± 0.0611	0.9046 ± 0.0679	0.9135 ± 0.0529	0.9192 ± 0.0587	0.9229 ± 0.0587	0.9329 ± 0.0483	0.9406 ± 0.0455
GMM Detector	RTVAE CORE	0.6254 ± 0.1768	0.6239 ± 0.1945	0.7115 ± 0.2103	0.7909 ± 0.1448	0.8015 ± 0.1859	0.8785 ± 0.1068	0.8708 ± 0.1613	0.8592 ± 0.1741	0.8806 ± 0.1319	0.8886 ± 0.1711	0.8965 ± 0.1568	0.9384 ± 0.0999	0.9525 ± 0.0963
XGBoost Detector	RTVAE CORE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
MLP Detector	RTVAE CORE	0.7154 ± 0.1666	0.7611 ± 0.1494	0.7946 ± 0.1380	0.8212 ± 0.1321	0.8346 ± 0.1303	0.8482 ± 0.1285	0.8566 ± 0.1287	0.8663 ± 0.1288	0.8716 ± 0.1281	0.8805 ± 0.1283	0.8827 ± 0.1279	0.8686 ± 0.1694	0.8765 ± 0.1724
Linear Detector	RTVAE CORE	0.7316 ± 0.1847	0.7867 ± 0.1483	0.8248 ± 0.1221	0.8543 ± 0.1017	0.8701 ± 0.0920	0.8782 ± 0.0880	0.8837 ± 0.0854	0.8929 ± 0.0784	0.9016 ± 0.0713	0.9079 ± 0.0675	0.9124 ± 0.0637	0.9249 ± 0.0563	0.9333 ± 0.0495
SDMetrics Column Shapes	RTVAE CORE	0.7868 ± 0.1625	0.7589 ± 0.1568	0.7371 ± 0.1496	0.7209 ± 0.1424	0.7087 ± 0.1367	0.7002 ± 0.1337	0.6911 ± 0.1349	0.6822 ± 0.1310	0.6737 ± 0.1285	0.6672 ± 0.1249	0.6642 ± 0.1238	0.6499 ± 0.1175	0.6456 ± 0.1167
SDMetrics Column Pair Trends	RTVAE CORE	0.8904 ± 0.0299	0.8473 ± 0.0148	0.8103 ± 0.0130	0.7791 ± 0.0196	0.7572 ± 0.0199	0.7411 ± 0.0179	0.7321 ± 0.0239	0.7240 ± 0.0273	0.7166 ± 0.0268	0.7162 ± 0.0292	0.7135 ± 0.0295	0.7147 ± 0.0323	0.7064 ± 0.0303
Memorized Fraction	RTVAE CORE	1.8793 ± 1.2976	3.1513 ± 2.5335	3.9947 ± 3.2344	4.5940 ± 3.6800	5.0160 ± 4.0943	5.3460 ± 4.3806	6.1580 ± 5.2271	6.6313 ± 5.7349	6.6833 ± 5.7886	6.9900 ± 6.2720	7.5200 ± 6.5393	9.0500 ± 8.1015	11.1747 ± 9.2541
New Row Synthesis	RTVAE CORE	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Training Time (s)	RTVAE CORE	184.7080 ± 70.9919	158.4527 ± 25.9206	162.0760 ± 24.6240	179.0973 ± 20.2988	169.8400 ± 27.3630	190.5380 ± 35.9449	202.8113 ± 67.4529	184.4987 ± 21.6364	196.1187 ± 39.0925	202.5187 ± 35.5388	198.9147 ± 31.8524	237.0693 ± 36.5772	235.2033 ± 53.0283
Generation Time (s)	RTVAE CORE	0.7153 ± 0.8745	0.4940 ± 0.7528	0.3160 ± 0.6928	0.5087 ± 0.8030	0.1613 ± 0.0439	0.4027 ± 0.6972	0.2287 ± 0.2915	0.3273 ± 0.5015	0.6440 ± 1.0242	0.4153 ± 0.9809	0.2993 ± 0.5211	0.4127 ± 0.6814	0.3747 ± 0.5701
Model Size (MB)	RTVAE CORE	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
Minority class share (readm)	RTVAE CORE	0.0851 ± 0.0184	0.0431 ± 0.0213	0.0157 ± 0.0164	0.0034 ± 0.0084	0.0005 ± 0.0021	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Readmission Collapsed	RTVAE CORE	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0667 ± 0.2860	0.4667 ± 0.5719	0.8667 ± 0.3897	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
TRTR AUROC	RTVAE CORE	0.6162 ± 0.0078	0.6162 ± 0.0078	0.6163 ± 0.0082	0.6146 ± 0.0175	0.6242 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR AUROC	RTVAE CORE	0.5631 ± 0.0478	0.5529 ± 0.0344	0.5319 ± 0.0373	0.5076 ± 0.1125	0.5383 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR Accuracy	RTVAE CORE	0.6881 ± 0.1728	0.8057 ± 0.1081	0.8701 ± 0.0112	0.8723 ± 0.0020	0.8727 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan

metric	generator	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10	Gen 15	Gen 19
TSTR Precision	RTVAE CORE	0.1309 ± 0.0780	0.1000 ± 0.0979	0.0138 ± 0.0599	0.0182 ± 0.1010	0.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR Recall	RTVAE CORE	0.3270 ± 0.2419	0.1109 ± 0.1524	0.0056 ± 0.0245	0.0003 ± 0.0017	0.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR F1	RTVAE CORE	0.1793 ± 0.1063	0.0870 ± 0.0992	0.0080 ± 0.0348	0.0006 ± 0.0034	0.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR FPR	RTVAE CORE	0.2593 ± 0.2317	0.0929 ± 0.1453	0.0038 ± 0.0164	0.0005 ± 0.0025	0.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR FNR	RTVAE CORE	0.6730 ± 0.2419	0.8891 ± 0.1524	0.9944 ± 0.0245	0.9997 ± 0.0017	1.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan
TSTR TNR	RTVAE CORE	0.7407 ± 0.2317	0.9071 ± 0.1453	0.9962 ± 0.0164	0.9995 ± 0.0025	1.0000 ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan	nan ± nan

2.4 (Large) Results tables from statistical analysis

Supplementary table 19Log-rank and stratified Cox comparisons of time-to-failure distributions, by dataset.

Omnibus and Holm-adjusted pairwise log-rank test statistics, and Cox proportional-hazards likelihood-ratio test statistics stratified on the (data seed, model seed) pair, for the full population and each leave-one-seed-out subset, for all four groups. Where a group includes a generator failing in 100% of its chains, a supplementary Cox model excluding that generator is reported alongside the primary model.

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full	Full population	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	72.20683	3	1.44E-15	
Full	Full population	Pairwise log-rank	ARF vs CTGAN	1.228111	1	0.267775	0.267775
Full	Full population	Pairwise log-rank	ARF vs DDPM	10.52841	1	0.001176	0.003527
Full	Full population	Pairwise log-rank	ARF vs RTVAE	28.18742	1	1.10E-07	4.40E-07
Full	Full population	Pairwise log-rank	CTGAN vs DDPM	5.414323	1	0.019972	0.039944

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full	Full population	Pairwise log-rank	CTGAN vs RTVAE	31.70108	1	1.80E-08	1.08E-07
Full	Full population	Pairwise log-rank	DDPM vs RTVAE	29.80184	1	4.79E-08	2.39E-07
Full	Full population	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	44.10887	3	1.43E-09	
Full	Full population	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	8.720539	2	0.012775	
Full	Ex data seed 1597	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	47.29316	3	3.01E-10	
Full	Ex data seed 1597	Pairwise log-rank	ARF vs CTGAN	0.009257	1	0.92335	0.92335
Full	Ex data seed 1597	Pairwise log-rank	ARF vs DDPM	4.42855	1	0.035343	0.090192
Full	Ex data seed 1597	Pairwise log-rank	ARF vs RTVAE	18.41728	1	1.77E-05	8.87E-05
Full	Ex data seed 1597	Pairwise log-rank	CTGAN vs DDPM	4.705636	1	0.030064	0.090192
Full	Ex data seed 1597	Pairwise log-rank	CTGAN vs RTVAE	20.54837	1	5.81E-06	3.49E-05
Full	Ex data seed 1597	Pairwise log-rank	DDPM vs RTVAE	18.41728	1	1.77E-05	8.87E-05
Full	Ex data seed 1597	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	27.30516	3	5.08E-06	
Full	Ex data seed 1597	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	5.04356	2	0.080317	
Full	Ex data seed 196418	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	53.94012	3	1.16E-11	
Full	Ex data seed 196418	Pairwise log-rank	ARF vs CTGAN	4.73642	1	0.02953	0.059061
Full	Ex data seed 196418	Pairwise log-rank	ARF vs DDPM	16.40866	1	5.11E-05	0.000153
Full	Ex data seed 196418	Pairwise log-rank	ARF vs RTVAE	19.48685	1	1.01E-05	4.05E-05
Full	Ex data seed 196418	Pairwise log-rank	CTGAN vs DDPM	4.10438	1	0.042772	0.059061
Full	Ex data seed 196418	Pairwise log-rank	CTGAN vs RTVAE	21.12089	1	4.31E-06	2.59E-05

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full	Ex data seed 196418	Pairwise log-rank	DDPM vs RTVAE	21.12089	1	4.31E-06	2.59E-05
Full	Ex data seed 196418	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	38.57568	3	2.13E-08	
Full	Ex data seed 196418	Stratified Cox LRT (Ex ARF, RTVAE)	CTGAN vs DDPM	2.092993	1	0.147976	
Full	Ex data seed 14930352	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	44.55995	3	1.15E-09	
Full	Ex data seed 14930352	Pairwise log-rank	ARF vs CTGAN	0.455884	1	0.499554	0.499554
Full	Ex data seed 14930352	Pairwise log-rank	ARF vs DDPM	3.737477	1	0.053205	0.159614
Full	Ex data seed 14930352	Pairwise log-rank	ARF vs RTVAE	17.73629	1	2.54E-05	0.000101
Full	Ex data seed 14930352	Pairwise log-rank	CTGAN vs DDPM	2.084038	1	0.148846	0.297692
Full	Ex data seed 14930352	Pairwise log-rank	CTGAN vs RTVAE	20.35481	1	6.43E-06	3.86E-05
Full	Ex data seed 14930352	Pairwise log-rank	DDPM vs RTVAE	18.89631	1	1.38E-05	6.90E-05
Full	Ex data seed 14930352	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	26.24219	3	8.49E-06	
Full	Ex data seed 14930352	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	3.977598	2	0.13686	
Full	Ex model seed 987	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	57.5653	3	1.95E-12	
Full	Ex model seed 987	Pairwise log-rank	ARF vs CTGAN	0.329825	1	0.565762	0.565762
Full	Ex model seed 987	Pairwise log-rank	ARF vs DDPM	8.428848	1	0.003693	0.011079
Full	Ex model seed 987	Pairwise log-rank	ARF vs RTVAE	23.56453	1	1.21E-06	6.04E-06
Full	Ex model seed 987	Pairwise log-rank	CTGAN vs DDPM	5.909379	1	0.01506	0.030121
Full	Ex model seed 987	Pairwise log-rank	CTGAN vs RTVAE	25.79338	1	3.80E-07	2.28E-06
Full	Ex model seed 987	Pairwise log-rank	DDPM vs RTVAE	23.56453	1	1.21E-06	6.04E-06

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full	Ex model seed 987	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	36.81105	3	5.05E-08	
Full	Ex model seed 987	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	9.338232	2	0.009381	
Full	Ex model seed 28657	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	59.32615	3	8.19E-13	
Full	Ex model seed 28657	Pairwise log-rank	ARF vs CTGAN	2.061075	1	0.151104	0.151104
Full	Ex model seed 28657	Pairwise log-rank	ARF vs DDPM	11.56162	1	0.000673	0.00202
Full	Ex model seed 28657	Pairwise log-rank	ARF vs RTVAE	22.12218	1	2.56E-06	1.02E-05
Full	Ex model seed 28657	Pairwise log-rank	CTGAN vs DDPM	3.735444	1	0.05327	0.106539
Full	Ex model seed 28657	Pairwise log-rank	CTGAN vs RTVAE	24.55927	1	7.21E-07	4.32E-06
Full	Ex model seed 28657	Pairwise log-rank	DDPM vs RTVAE	24.55927	1	7.21E-07	4.32E-06
Full	Ex model seed 28657	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	41.59298	3	4.89E-09	
Full	Ex model seed 28657	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	8.32192	2	0.015593	
Full	Ex model seed 5702887	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	55.76259	3	4.72E-12	
Full	Ex model seed 5702887	Pairwise log-rank	ARF vs CTGAN	1.555673	1	0.2123	0.2123
Full	Ex model seed 5702887	Pairwise log-rank	ARF vs DDPM	6.647843	1	0.009928	0.029783
Full	Ex model seed 5702887	Pairwise log-rank	ARF vs RTVAE	22.12218	1	2.56E-06	1.02E-05
Full	Ex model seed 5702887	Pairwise log-rank	CTGAN vs DDPM	3.040021	1	0.081235	0.16247
Full	Ex model seed 5702887	Pairwise log-rank	CTGAN vs RTVAE	24.55927	1	7.21E-07	4.32E-06
Full	Ex model seed 5702887	Pairwise log-rank	DDPM vs RTVAE	23.22807	1	1.44E-06	7.19E-06
Full	Ex model seed 5702887	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	33.31691	3	2.76E-07	

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full	Ex model seed 5702887	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	5.835925	2	0.054044	
Full	Ex model seed 24157817	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	58.97672	3	9.72E-13	
Full	Ex model seed 24157817	Pairwise log-rank	ARF vs CTGAN	0.072776	1	0.787338	0.787338
Full	Ex model seed 24157817	Pairwise log-rank	ARF vs DDPM	7.341364	1	0.006739	0.020216
Full	Ex model seed 24157817	Pairwise log-rank	ARF vs RTVAE	23.0124	1	1.61E-06	8.05E-06
Full	Ex model seed 24157817	Pairwise log-rank	CTGAN vs DDPM	7.320461	1	0.006817	0.020216
Full	Ex model seed 24157817	Pairwise log-rank	CTGAN vs RTVAE	25.06192	1	5.55E-07	3.33E-06
Full	Ex model seed 24157817	Pairwise log-rank	DDPM vs RTVAE	23.0124	1	1.61E-06	8.05E-06
Full	Ex model seed 24157817	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	34.75756	3	1.37E-07	
Full	Ex model seed 24157817	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	7.280582	2	0.026245	
Full	Ex model seed 39088169	Omnibus log-rank	ARF vs CTGAN vs DDPM vs RTVAE	56.89712	3	2.70E-12	
Full	Ex model seed 39088169	Pairwise log-rank	ARF vs CTGAN	2.185232	1	0.13934	0.235153
Full	Ex model seed 39088169	Pairwise log-rank	ARF vs DDPM	8.479793	1	0.003591	0.010773
Full	Ex model seed 39088169	Pairwise log-rank	ARF vs RTVAE	21.34385	1	3.84E-06	1.54E-05
Full	Ex model seed 39088169	Pairwise log-rank	CTGAN vs DDPM	2.449311	1	0.117576	0.235153
Full	Ex model seed 39088169	Pairwise log-rank	CTGAN vs RTVAE	25.06192	1	5.55E-07	3.33E-06
Full	Ex model seed 39088169	Pairwise log-rank	DDPM vs RTVAE	23.0124	1	1.61E-06	8.05E-06
Full	Ex model seed 39088169	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM vs RTVAE	33.21927	3	2.90E-07	
Full	Ex model seed 39088169	Stratified Cox LRT (Ex RTVAE)	ARF vs CTGAN vs DDPM	5.737436	2	0.056772	

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core	Full population	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	74.22716	3	5.31E-16	
Core	Full population	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.421525	1	0.516177	1
Core	Full population	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.080904	1	0.776076	1
Core	Full population	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	31.97176	1	1.56E-08	9.39E-08
Core	Full population	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.383366	1	0.535807	1
Core	Full population	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	31.97176	1	1.56E-08	9.39E-08
Core	Full population	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	27.12596	1	1.91E-07	7.62E-07
Core	Full population	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	37.45137	3	3.69E-08	
Core	Full population	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.030291	2	0.36235	
Core	Ex data seed 1597	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	51.23527	3	4.36E-11	
Core	Ex data seed 1597	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.297737	1	0.585304	1
Core	Ex data seed 1597	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.050365	1	0.822429	1
Core	Ex data seed 1597	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	20.72312	1	5.31E-06	3.18E-05
Core	Ex data seed 1597	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.552108	1	0.457458	1
Core	Ex data seed 1597	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	20.72312	1	5.31E-06	3.18E-05
Core	Ex data seed 1597	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	18.7932	1	1.46E-05	5.83E-05
Core	Ex data seed 1597	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	29.27933	3	1.96E-06	
Core	Ex data seed 1597	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	1.553439	2	0.459912	
Core	Ex data seed 196418	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	48.70849	3	1.50E-10	

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core	Ex data seed 196418	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	2.813052	1	0.0935	0.280501
Core	Ex data seed 196418	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.012569	1	0.910735	0.910735
Core	Ex data seed 196418	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	21.22554	1	4.08E-06	2.45E-05
Core	Ex data seed 196418	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	1.143019	1	0.285015	0.570031
Core	Ex data seed 196418	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	21.22554	1	4.08E-06	2.45E-05
Core	Ex data seed 196418	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	17.13473	1	3.48E-05	0.000139
Core	Ex data seed 196418	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	25.39442	3	1.28E-05	
Core	Ex data seed 196418	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	3.093966	2	0.212889	
Core	Ex data seed 14930352	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	47.21432	3	3.13E-10	
Core	Ex data seed 14930352	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.217717	1	0.640785	1
Core	Ex data seed 14930352	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.226692	1	0.633988	1
Core	Ex data seed 14930352	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	20.31951	1	6.55E-06	3.93E-05
Core	Ex data seed 14930352	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.058225	1	0.809324	1
Core	Ex data seed 14930352	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	20.31951	1	6.55E-06	3.93E-05
Core	Ex data seed 14930352	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	17.24542	1	3.28E-05	0.000131
Core	Ex data seed 14930352	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	22.84995	3	4.34E-05	
Core	Ex data seed 14930352	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.539879	2	0.763426	
Core	Ex model seed 987	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	64.59823	3	6.11E-14	
Core	Ex model seed 987	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.629094	1	0.427688	0.672799

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core	Ex model seed 987	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.924105	1	0.336399	0.672799
Core	Ex model seed 987	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	25.44631	1	4.55E-07	2.73E-06
Core	Ex model seed 987	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	2.226456	1	0.135665	0.406994
Core	Ex model seed 987	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	25.44631	1	4.55E-07	2.73E-06
Core	Ex model seed 987	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	25.44631	1	4.55E-07	2.73E-06
Core	Ex model seed 987	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	35.36148	3	1.02E-07	
Core	Ex model seed 987	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.09042	2	0.351618	
Core	Ex model seed 28657	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	57.78909	3	1.74E-12	
Core	Ex model seed 28657	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.010165	1	0.919693	1
Core	Ex model seed 28657	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.595383	1	0.440345	1
Core	Ex model seed 28657	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	25.44631	1	4.55E-07	2.73E-06
Core	Ex model seed 28657	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.263082	1	0.60801	1
Core	Ex model seed 28657	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	25.44631	1	4.55E-07	2.73E-06
Core	Ex model seed 28657	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	21.42041	1	3.69E-06	1.48E-05
Core	Ex model seed 28657	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	27.66517	3	4.27E-06	
Core	Ex model seed 28657	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.151952	2	0.926839	
Core	Ex model seed 5702887	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	58.47084	3	1.25E-12	
Core	Ex model seed 5702887	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.466811	1	0.494458	1
Core	Ex model seed 5702887	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.262193	1	0.608618	1

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core	Ex model seed 5702887	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	25.18935	1	5.20E-07	3.12E-06
Core	Ex model seed 5702887	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.883595	1	0.347219	1
Core	Ex model seed 5702887	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	25.18935	1	5.20E-07	3.12E-06
Core	Ex model seed 5702887	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	20.20669	1	6.95E-06	2.78E-05
Core	Ex model seed 5702887	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	28.51666	3	2.83E-06	
Core	Ex model seed 5702887	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	1.008916	2	0.603833	
Core	Ex model seed 24157817	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	58.38788	3	1.30E-12	
Core	Ex model seed 24157817	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	1.02594	1	0.311114	0.933342
Core	Ex model seed 24157817	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.481491	1	0.487748	0.975495
Core	Ex model seed 24157817	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	25.18935	1	5.20E-07	3.12E-06
Core	Ex model seed 24157817	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.006689	1	0.934815	0.975495
Core	Ex model seed 24157817	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	25.18935	1	5.20E-07	3.12E-06
Core	Ex model seed 24157817	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	20.20669	1	6.95E-06	2.78E-05
Core	Ex model seed 24157817	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	32.01931	3	5.18E-07	
Core	Ex model seed 24157817	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	4.48689	2	0.106092	
Core	Ex model seed 39088169	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	56.23417	3	3.74E-12	
Core	Ex model seed 39088169	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.306184	1	0.580031	1
Core	Ex model seed 39088169	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.206555	1	0.649481	1
Core	Ex model seed 39088169	Pairwise log-rank	ARF_CORE vs RTVAE_CORE	25.01546	1	5.69E-07	3.41E-06

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core	Ex model seed 39088169	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.00856	1	0.926283	1
Core	Ex model seed 39088169	Pairwise log-rank	CTGAN_CORE vs RTVAE_CORE	25.01546	1	5.69E-07	3.41E-06
Core	Ex model seed 39088169	Pairwise log-rank	DDPM_CORE vs RTVAE_CORE	19.87406	1	8.27E-06	3.31E-05
Core	Ex model seed 39088169	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE vs RTVAE_CORE	31.77085	3	5.85E-07	
Core	Ex model seed 39088169	Stratified Cox LRT (Ex RTVAE_CORE)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	4.23753	2	0.12018	
Full, Ex RTVAE	Full population	Omnibus log-rank	ARF vs CTGAN vs DDPM	11.14899	2	0.003793	
Full, Ex RTVAE	Full population	Pairwise log-rank	ARF vs CTGAN	1.228111	1	0.267775	0.267775
Full, Ex RTVAE	Full population	Pairwise log-rank	ARF vs DDPM	10.52841	1	0.001176	0.003527
Full, Ex RTVAE	Full population	Pairwise log-rank	CTGAN vs DDPM	5.414323	1	0.019972	0.039944
Full, Ex RTVAE	Full population	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	8.720539	2	0.012775	
Full, Ex RTVAE	Ex data seed 1597	Omnibus log-rank	ARF vs CTGAN vs DDPM	5.809851	2	0.054753	
Full, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	ARF vs CTGAN	0.009257	1	0.92335	0.92335
Full, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	ARF vs DDPM	4.42855	1	0.035343	0.090192
Full, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	CTGAN vs DDPM	4.705636	1	0.030064	0.090192
Full, Ex RTVAE	Ex data seed 1597	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	5.04356	2	0.080317	
Full, Ex RTVAE	Ex data seed 196418	Omnibus log-rank	ARF vs CTGAN vs DDPM	17.22918	2	0.000181	
Full, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	ARF vs CTGAN	4.73642	1	0.02953	0.059061
Full, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	ARF vs DDPM	16.40866	1	5.11E-05	0.000153
Full, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	CTGAN vs DDPM	4.10438	1	0.042772	0.059061

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full, Ex RTVAE	Ex data seed 196418	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	10.84979	2	0.004406	
Full, Ex RTVAE	Ex data seed 196418	Stratified Cox LRT (Ex ARF)	CTGAN vs DDPM	2.092993	1	0.147976	
Full, Ex RTVAE	Ex data seed 14930352	Omnibus log-rank	ARF vs CTGAN vs DDPM	4.232197	2	0.120501	
Full, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	ARF vs CTGAN	0.455884	1	0.499554	0.499554
Full, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	ARF vs DDPM	3.737477	1	0.053205	0.159614
Full, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	CTGAN vs DDPM	2.084038	1	0.148846	0.297692
Full, Ex RTVAE	Ex data seed 14930352	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	3.977598	2	0.13686	
Full, Ex RTVAE	Ex model seed 987	Omnibus log-rank	ARF vs CTGAN vs DDPM	9.193256	2	0.010086	
Full, Ex RTVAE	Ex model seed 987	Pairwise log-rank	ARF vs CTGAN	0.329825	1	0.565762	0.565762
Full, Ex RTVAE	Ex model seed 987	Pairwise log-rank	ARF vs DDPM	8.428848	1	0.003693	0.011079
Full, Ex RTVAE	Ex model seed 987	Pairwise log-rank	CTGAN vs DDPM	5.909379	1	0.01506	0.030121
Full, Ex RTVAE	Ex model seed 987	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	9.338232	2	0.009381	
Full, Ex RTVAE	Ex model seed 28657	Omnibus log-rank	ARF vs CTGAN vs DDPM	11.56807	2	0.003076	
Full, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	ARF vs CTGAN	2.061075	1	0.151104	0.151104
Full, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	ARF vs DDPM	11.56162	1	0.000673	0.00202
Full, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	CTGAN vs DDPM	3.735444	1	0.05327	0.106539
Full, Ex RTVAE	Ex model seed 28657	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	8.32192	2	0.015593	
Full, Ex RTVAE	Ex model seed 5702887	Omnibus log-rank	ARF vs CTGAN vs DDPM	7.823765	2	0.020003	
Full, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	ARF vs CTGAN	1.555673	1	0.2123	0.2123

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Full, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	ARF vs DDPM	6.647843	1	0.009928	0.029783
Full, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	CTGAN vs DDPM	3.040021	1	0.081235	0.16247
Full, Ex RTVAE	Ex model seed 5702887	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	5.835925	2	0.054044	
Full, Ex RTVAE	Ex model seed 24157817	Omnibus log-rank	ARF vs CTGAN vs DDPM	9.0193	2	0.011002	
Full, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	ARF vs CTGAN	0.072776	1	0.787338	0.787338
Full, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	ARF vs DDPM	7.341364	1	0.006739	0.020216
Full, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	CTGAN vs DDPM	7.320461	1	0.006817	0.020216
Full, Ex RTVAE	Ex model seed 24157817	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	7.280582	2	0.026245	
Full, Ex RTVAE	Ex model seed 39088169	Omnibus log-rank	ARF vs CTGAN vs DDPM	8.955762	2	0.011357	
Full, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	ARF vs CTGAN	2.185232	1	0.13934	0.235153
Full, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	ARF vs DDPM	8.479793	1	0.003591	0.010773
Full, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	CTGAN vs DDPM	2.449311	1	0.117576	0.235153
Full, Ex RTVAE	Ex model seed 39088169	Stratified Cox LRT (all generators)	ARF vs CTGAN vs DDPM	5.737436	2	0.056772	
Core, Ex RTVAE	Full population	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.632338	2	0.728936	
Core, Ex RTVAE	Full population	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.421525	1	0.516177	1
Core, Ex RTVAE	Full population	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.080904	1	0.776076	1
Core, Ex RTVAE	Full population	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.383366	1	0.535807	1
Core, Ex RTVAE	Full population	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.030291	2	0.36235	
Core, Ex RTVAE	Ex data seed 1597	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.662914	2	0.717877	

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.297737	1	0.585304	1
Core, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.050365	1	0.822429	1
Core, Ex RTVAE	Ex data seed 1597	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.552108	1	0.457458	1
Core, Ex RTVAE	Ex data seed 1597	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	1.553439	2	0.459912	
Core, Ex RTVAE	Ex data seed 196418	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.570353	2	0.276602	
Core, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	2.813052	1	0.0935	0.280501
Core, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.012569	1	0.910735	0.910735
Core, Ex RTVAE	Ex data seed 196418	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	1.143019	1	0.285015	0.570031
Core, Ex RTVAE	Ex data seed 196418	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	3.093966	2	0.212889	
Core, Ex RTVAE	Ex data seed 14930352	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.268881	2	0.874205	
Core, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.217717	1	0.640785	1
Core, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.226692	1	0.633988	1
Core, Ex RTVAE	Ex data seed 14930352	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.058225	1	0.809324	1
Core, Ex RTVAE	Ex data seed 14930352	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.539879	2	0.763426	
Core, Ex RTVAE	Ex model seed 987	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.658914	2	0.264621	
Core, Ex RTVAE	Ex model seed 987	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.629094	1	0.427688	0.672799
Core, Ex RTVAE	Ex model seed 987	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.924105	1	0.336399	0.672799
Core, Ex RTVAE	Ex model seed 987	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	2.226456	1	0.135665	0.406994
Core, Ex RTVAE	Ex model seed 987	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	2.09042	2	0.351618	

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core, Ex RTVAE	Ex model seed 28657	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.593462	2	0.743244	
Core, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.010165	1	0.919693	1
Core, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.595383	1	0.440345	1
Core, Ex RTVAE	Ex model seed 28657	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.263082	1	0.60801	1
Core, Ex RTVAE	Ex model seed 28657	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.151952	2	0.926839	
Core, Ex RTVAE	Ex model seed 5702887	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	1.142741	2	0.564751	
Core, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.466811	1	0.494458	1
Core, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.262193	1	0.608618	1
Core, Ex RTVAE	Ex model seed 5702887	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.883595	1	0.347219	1
Core, Ex RTVAE	Ex model seed 5702887	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	1.008916	2	0.603833	
Core, Ex RTVAE	Ex model seed 24157817	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.931098	2	0.62779	
Core, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	1.02594	1	0.311114	0.933342
Core, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.481491	1	0.487748	0.975495
Core, Ex RTVAE	Ex model seed 24157817	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.006689	1	0.934815	0.975495
Core, Ex RTVAE	Ex model seed 24157817	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	4.48689	2	0.106092	
Core, Ex RTVAE	Ex model seed 39088169	Omnibus log-rank	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	0.335708	2	0.845477	
Core, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	ARF_CORE vs CTGAN_CORE	0.306184	1	0.580031	1
Core, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	ARF_CORE vs DDPM_CORE	0.206555	1	0.649481	1
Core, Ex RTVAE	Ex model seed 39088169	Pairwise log-rank	CTGAN_CORE vs DDPM_CORE	0.00856	1	0.926283	1

Group	Subset	Test	Comparison	Statistic	df	p	p_holm
Core, Ex RTVAE	Ex model seed 39088169	Stratified Cox LRT (all generators)	ARF_CORE vs CTGAN_CORE vs DDPM_CORE	4.23753	2	0.12018	