# Package 'syntheticdata'

April 2, 2026

**Title** Synthetic Clinical Data Generation and Privacy-Preserving Validation

**Version** 0.1.0

**Description** Generates synthetic clinical datasets that preserve statistical properties while reducing re-identification risk. Implements Gaussian copula simulation, bootstrap with noise injection, and Laplace noise perturbation, with built-in utility and privacy validation metrics. Useful for privacy-aware data sharing in multi-site clinical research. Validates synthetic data quality via distributional similarity (Kolmogorov-Smirnov), discriminative accuracy (real-vs-synthetic classifier), and nearest-neighbor privacy ratio. Methods described in Jordon et al. (2022) <doi:10.48550/arXiv.2205.03257> and Snoke et al. (2018) <doi:10.1111/rssa.12358>.

**License** MIT + file LICENSE

**URL** https://github.com/CuiweiG/syntheticdata

**BugReports** https://github.com/CuiweiG/syntheticdata/issues

**Depends** R (>= 4.1.0)

**Imports** cli (>= 3.4.0), dplyr (>= 1.1.0), stats, tibble (>= 3.1.0)

**Suggests** knitr, rmarkdown, testthat (>= 3.0.0)

**VignetteBuilder** knitr

**Config/testthat/edition** 3

**Language** en-US

**Encoding** UTF-8

**RoxygenNote** 7.3.3

**NeedsCompilation** no

**Author** Cuiwei Gao [aut, cre, cph]

**Maintainer** Cuiwei Gao <48gaocuiwei@gmail.com>

**Repository** CRAN

**Date/Publication** 2026-04-02 20:20:02 UTC

# Contents



---

`compare_methods` *Compare multiple synthesis methods*

---

## Description

Runs all three synthesis methods on the same data and returns a comparative validation table.

## Usage

```
compare_methods(data, n = nrow(data), seed = NULL)
```

## Arguments

| | |
|---|---|
| `data` | A data frame of real data. |
| `n` | Number of synthetic records. Default: same as input. |
| `seed` | Random seed passed to `synthesize()`. |

## Value

A `method_comparison` object (tibble) with columns: `method`, `metric`, `value`, `interpretation`.

## References

Jordon J, et al. (2022). Synthetic Data – what, why and how? *arXiv preprint* arXiv:2205.03257. doi:10.48550/arXiv.2205.03257

## Examples

```
set.seed(42)
real <- data.frame(x = rnorm(100), y = rnorm(100))
compare_methods(real, seed = 42)
```

`model_fidelity` *Downstream model fidelity test*

## Description

Trains a predictive model on synthetic data and evaluates it on real data. Compares to a model trained on real data (gold standard). Measures whether synthetic data preserves predictive signal.

## Usage

```
model_fidelity(x, outcome, predictors = NULL)
```

## Arguments

| | |
|---|---|
| `x` | A `synthetic_data` object from `synthesize()`. |
| `outcome` | Character. Name of the outcome column. |
| `predictors` | Character vector (optional). Predictor columns. Default: all other numeric columns. |

## Details

The real-data baseline uses in-sample evaluation (train and test on the same real data) to provide an upper bound on achievable performance. The synthetic-data model is also evaluated on real data, so the comparison reflects how well the synthetic data preserves predictive signal.

## Value

A tibble with columns: `train_data`, `metric`, `value`. For binary outcomes the metric is AUC; for continuous outcomes it is R-squared.

## References

Jordon J, et al. (2022). Synthetic Data – what, why and how? *arXiv preprint* arXiv:2205.03257. doi:10.48550/arXiv.2205.03257

## Examples

```
set.seed(42)
real <- data.frame(
    x1 = rnorm(200), x2 = rnorm(200),
    y = rbinom(200, 1, 0.3))
syn <- synthesize(real, seed = 42)
model_fidelity(syn, outcome = "y")
```

## privacy_risk *Compute privacy risk metrics*

### Description

Evaluates re-identification risk of synthetic data through multiple privacy metrics: nearest-neighbor distance ratio, membership inference accuracy, and attribute disclosure risk.

### Usage

```
privacy_risk(x, sensitive_cols = NULL)
```

### Arguments

| | |
|---|---|
| x | A synthetic_data object from synthesize(). |
| sensitive_cols | Character vector (optional). Columns considered sensitive for attribute disclosure assessment. |

### Value

A privacy_assessment object (tibble) with columns: metric, value, risk_level.

### References

Snoke J, et al. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society A*, 181(3):663–688. doi:10.1111/rssa.12358

### Examples

```
set.seed(42)
real <- data.frame(age = rnorm(100, 65, 10),
                   sbp = rnorm(100, 130, 20))
syn <- synthesize(real, seed = 42)
privacy_risk(syn)
```

## synthesize *Generate synthetic data from a real dataset*

### Description

Creates a synthetic version of the input data that preserves marginal distributions and pairwise correlations while adding controlled noise for privacy protection.

## Usage

```
synthesize(
  data,
  method = c("parametric", "bootstrap", "noise"),
  n = nrow(data),
  noise_level = 0.1,
  seed = NULL
)
```

## Arguments

| | |
|---|---|
| `data` | A data frame of real clinical data. |
| `method` | Synthesis method:<br>• `"parametric"` (default): fits Gaussian copula to continuous variables, multinomial to categorical. Fast, interpretable.<br>• `"bootstrap"`: nonparametric resampling with optional noise.<br>• `"noise"`: adds calibrated Laplace noise to each variable (differential privacy inspired). |
| `n` | Number of synthetic records. Default: same as input. |
| `noise_level` | For `method = "noise"`: scale of Laplace noise relative to variable SD. Default 0.1. |
| `seed` | Random seed for reproducibility. If non-NULL, the global RNG state is saved before and restored after synthesis so that calling code is not affected. |

## Details

The parametric method uses a Gaussian copula approach: marginal distributions are estimated empirically and the joint dependence structure is captured via the correlation matrix of normal scores. This preserves both marginal shapes and pairwise associations while generating genuinely new observations.

## Value

A `synthetic_data` object (list) with components: `$synthetic` (tibble of synthetic records), `$real` (tibble of the original data, retained for downstream validation), `$method`, `$n_original`, `$n_synthetic`, `$variables`.

## References

Jordon J, et al. (2022). Synthetic Data – what, why and how? *arXiv preprint* arXiv:2205.03257. doi:10.48550/arXiv.2205.03257

## Examples

```
set.seed(42)
real <- data.frame(
  age = rnorm(200, 65, 10),
```

```
  sbp = rnorm(200, 130, 20),
  sex = sample(c("M", "F"), 200, replace = TRUE),
  outcome = rbinom(200, 1, 0.3)
)
syn <- synthesize(real, method = "parametric", seed = 42)
syn
```

---

`validate_synthetic` *Validate synthetic data quality*

---

## Description

Computes utility and privacy metrics comparing synthetic data to the original real dataset.

## Usage

```
validate_synthetic(
  x,
  metrics = c("distributional", "correlation", "discriminative", "privacy")
)
```

## Arguments

| | |
|---|---|
| `x` | A `synthetic_data` object from `synthesize()`. |
| `metrics` | Character vector of metrics:<br>• `"distributional"`: KS statistic per numeric variable.<br>• `"correlation"`: Frobenius norm of correlation difference.<br>• `"discriminative"`: AUC of real-vs-synthetic classifier.<br>• `"privacy"`: nearest-neighbor distance ratio. |

## Details

Utility metrics assess how well the synthetic data preserves statistical properties. Privacy metrics assess the risk of re-identification.

Discriminative accuracy near 0.5 means the synthetic data is indistinguishable from real data. Privacy ratio > 1 means synthetic records are not closer to real records than real records are to each other.

## Value

A `synthetic_validation` object (tibble) with columns: `metric`, `value`, `interpretation`.

## References

Snoke J, et al. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society A*, 181(3):663–688. doi:10.1111/rssa.12358

## Examples

```
set.seed(42)
real <- data.frame(age = rnorm(100, 65, 10), sbp = rnorm(100, 130, 20))
syn <- synthesize(real, seed = 42)
validate_synthetic(syn)
```

# Index