

Efficient generation of virtual populations using machine learning in R

Alina Melnikova¹, Kirill Zhudenko^{1,2}, Leonid Stolbov¹, Victor Sokolov^{1,3}

1 - M&S Decisions, Dubai, United Arab Emirates, 2 - Research Center of Model-Informed Drug Development, Sechenov First Moscow State Medical University, Moscow, Russia, 3 - Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences (INM RAS), Moscow, Russia



Introduction

- Virtual population generation is a crucial component of quantitative pharmacology and dose optimization [1]. Bootstrap and copula-based methods are traditional approaches for generating subject-level covariates [2-4]. However, they have several known limitations: bootstrap lacks variability; copula methods limited by dimensionality-driven quality loss, preprocessing needs, and sensitivity to parametric assumptions [5,6].
- To overcome these issues, machine learning approaches are being developed for synthetic data generation [7,8].
- However, these emerging approaches require specific refinement for use in quantitative pharmacology applications. At present, there is still a shortage of such tools - publicly available and user-friendly.

Objectives

- Develop an R-based tool for data synthesis, diagnostics, and synthetic-vs-reference dataset comparison, tailored for clinical data, with native support for empirical distributions, including non-normal and imbalanced data.

Materials and Methods

Random forest-based synthetic data generation

- The tool for virtual population generation implemented into the **SimuRg** (see Poster I-102) R package is based on **random forest (RF)** method.

Table 1: Random forest advantages over Copula methods.

Nonlinear dependency modeling	Captures complex covariate interactions without assuming linearity or predefined functional forms
Nonparametric approach	No assumptions about marginal distributions and dependence structures
Mixed data support	Handles continuous and categorical variables naturally with minimal preprocessing

- RF is an ensemble machine learning method based on multiple decision trees. Each tree predicts the target variable from a set of covariates using different samples of the original dataset (Figure 1):

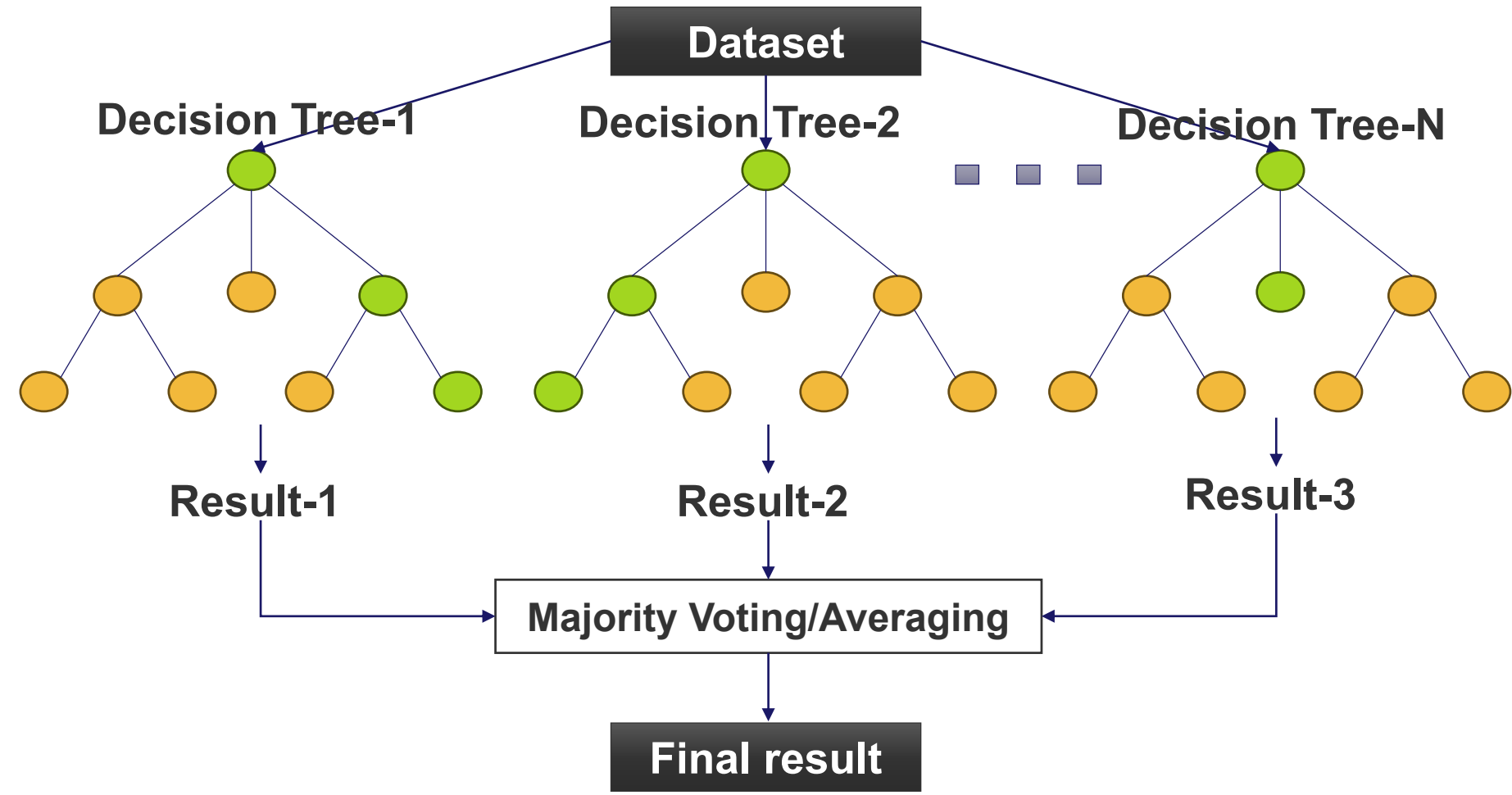


Figure 1: Scheme of RF method operation.

- In process of data generation variables are synthesized sequentially from conditional distributions $P(X_j|X_1, \dots, X_{j-1})$ estimated on the original dataset. For reconstruction of the conditional distribution RF method is selected (*synthpop* R package (ver. 1.9-2) [9]). Given different seeds, multiple replicas of synthetic covariate sets were generated (Figure 2).

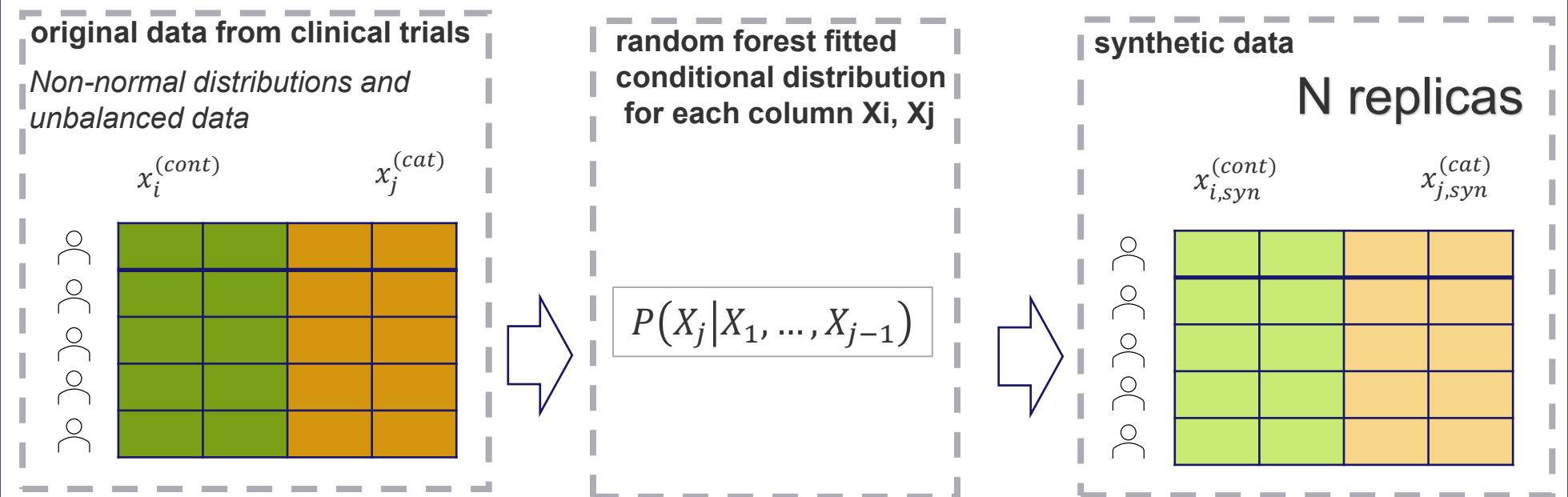


Figure 2: Scheme of synthetic data generation.

Reference datasets

- The algorithm of synthetic data generation was evaluated using 4 biomedical datasets: ACTG (**AIDS**) (467 subj., 2 cont. var., 5 cat. var.,[10]), **GBSG2** (686 subj., 5 cont. var., 4 cat. var.,[11]), **METABRIC** (1766 subj., 4 cont. var., 7 cat. var.,[12]), and **PBC** (280 subj., 6 cont. var., 7 cat. var.,[13]).

Synthetic data diagnostics and comparison

Table 2: List of diagnostic metrics used for synthetic data evaluation.

Similarity of continuous covariate distributions	- Kolmogorov-Smirnov test [14] - Comparative histograms of original and synthetic data
Similarity of categorical covariate distributions	- Jensen-Shannon divergence [15] - Comparative bar plots of original and synthetic data
Multivariate distribution equivalence	- Uniform Manifold Approximation and Projection (UMAP) [16] - The mean and the maximum absolute Kendall correlation coefficient difference (ΔCC) across all continuous-variable pairs
Data disclosure risk	Exact duplicates check

The complete virtual population generation and diagnostic workflow is implemented in the **SimuRg** R package (function *sg_vpop_est*).



Results

Univariate analysis of the data

- In all tested datasets continuous variables showed high agreement between original and synthetic data: mean Kolmogorov–Smirnov p-values ranged from 0.83 to 0.85, indicating no statistically significant differences between empirical and synthetic distributions (Table 3).
- Histograms further confirmed these findings, demonstrating close agreement in overall distribution shape (see the example for METABRIC data in Figure 3).

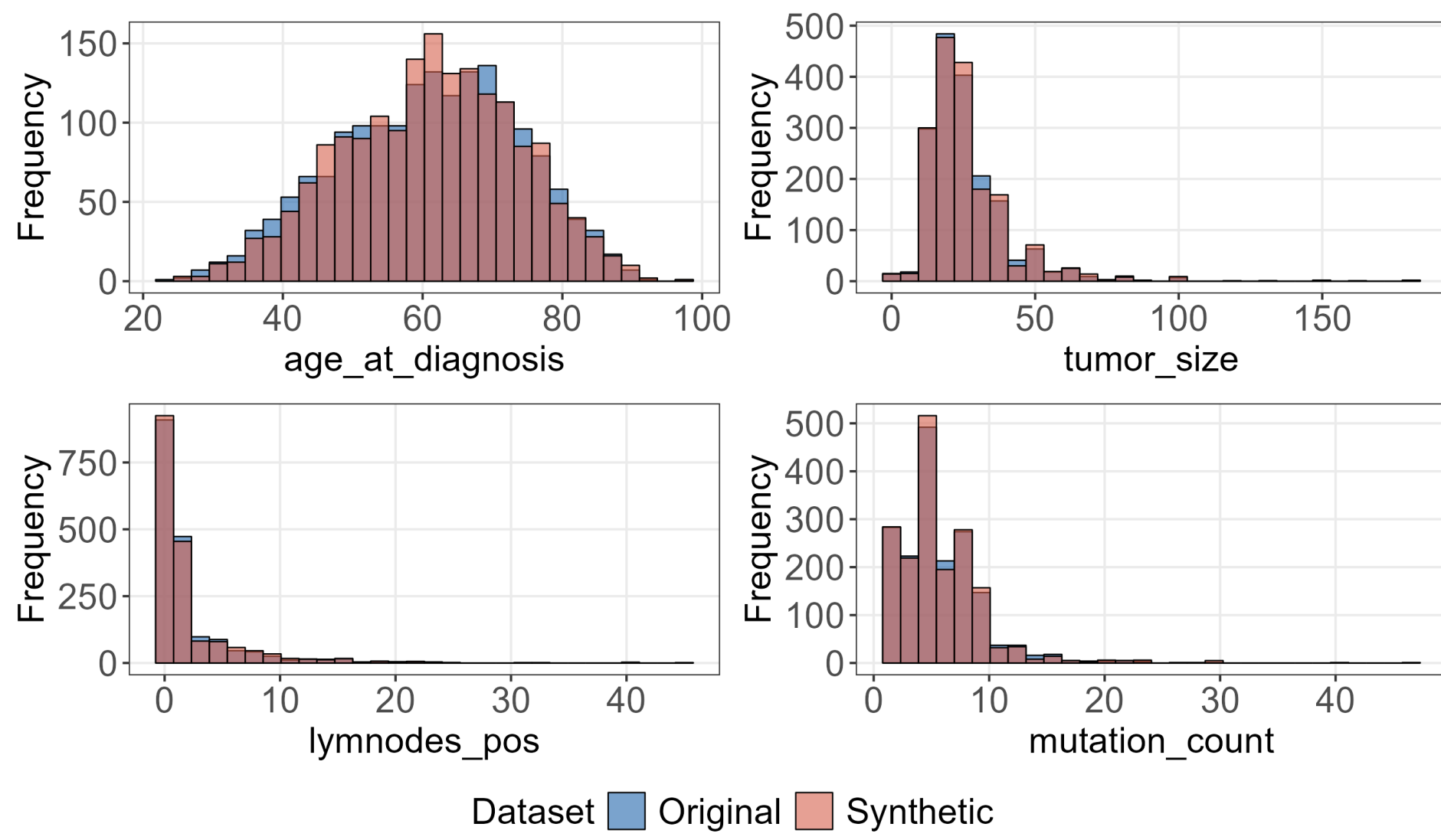


Figure 3: Histograms of the continuous variable distributions (METABRIC data, average seed).

- Categorical variables also demonstrated high similarity between original and synthetic datasets. Mean Jensen–Shannon divergence values are less than 0.01 for all datasets, indicating near-identical empirical distributions (Table 3). Data distribution histograms (Figure 4, example for METABRIC data) also demonstrated strong agreement across all categorical variable distributions.

Dataset name	KS test. mean p-value [95% CI]	100*JSD mean [95% CI]	Mean max ΔCC [95% CI]	Mean average ΔCC [95% CI]
AIDS	0.83 [0.80;0.86]	0.032 [0.028;0.036]	0.028 [0.024;0.032]	0.028 [0.024;0.032]
GBSG2	0.84 [0.83;0.86]	0.130 [0.12;0.14]	0.058 [0.054;0.062]	0.024 [0.022;0.025]
METABRIC	0.85 [0.82;0.87]	0.074 [0.070;0.078]	0.041 [0.038;0.044]	0.018 [0.017;0.019]
PBC	0.84 [0.82;0.86]	0.320 [0.30;0.34]	0.100 [0.095;0.11]	0.038 [0.036;0.039]

Table 3: Results of overall Kolmogorov-Smirnov tests, Jensen-Shannon divergence and correlation preservation metrics for all reference datasets.

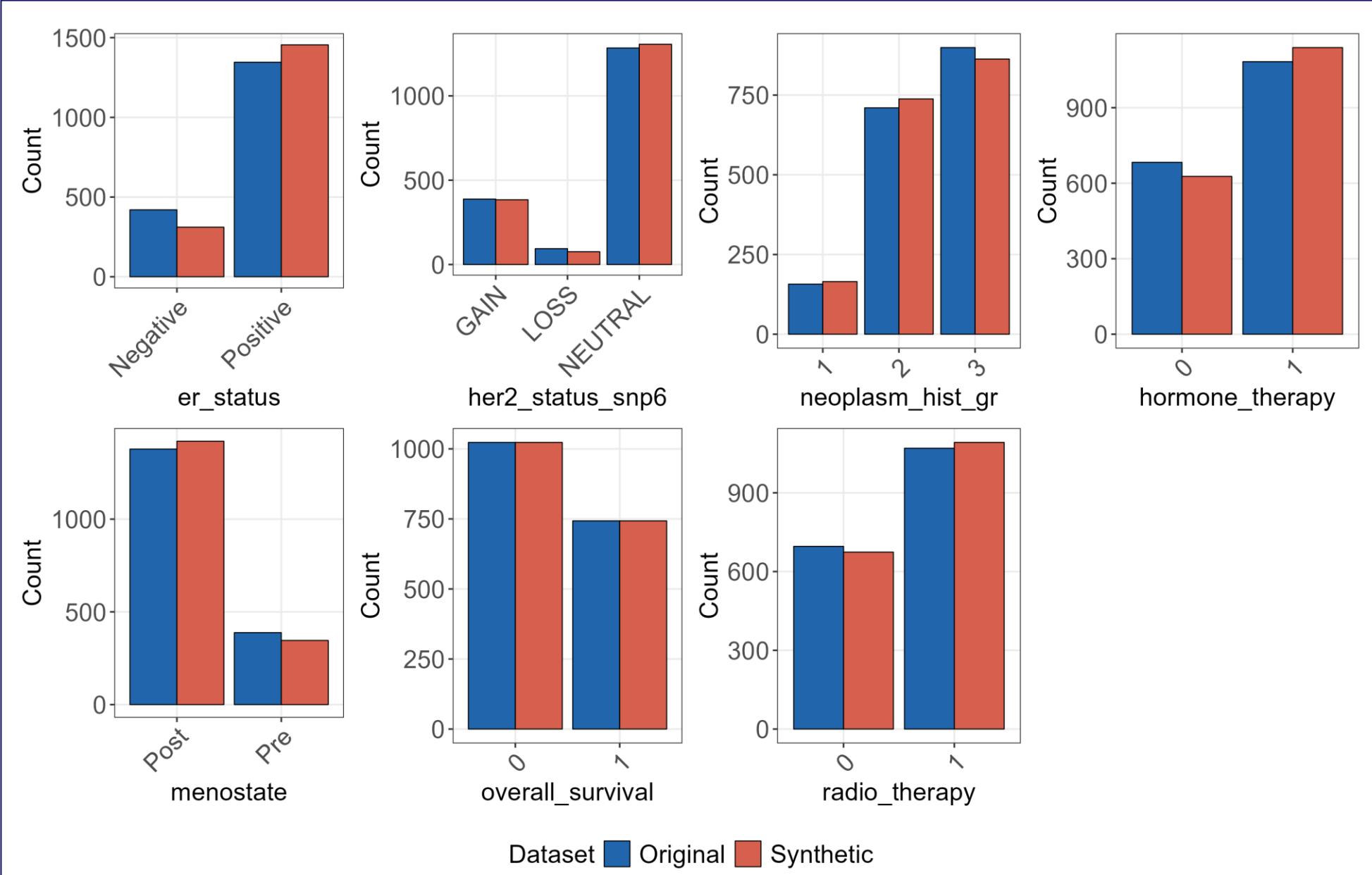


Figure 4: Histograms of categorical variable distributions (METABRIC, avr. seed).

Multivariate analysis of the data

- UMAP visualizations provided a comprehensive assessment of joint distribution across both continuous and categorical variables in investigated data. For all datasets UMAP plots demonstrated strong overlap between original and synthetic observation areas (Figure 5).

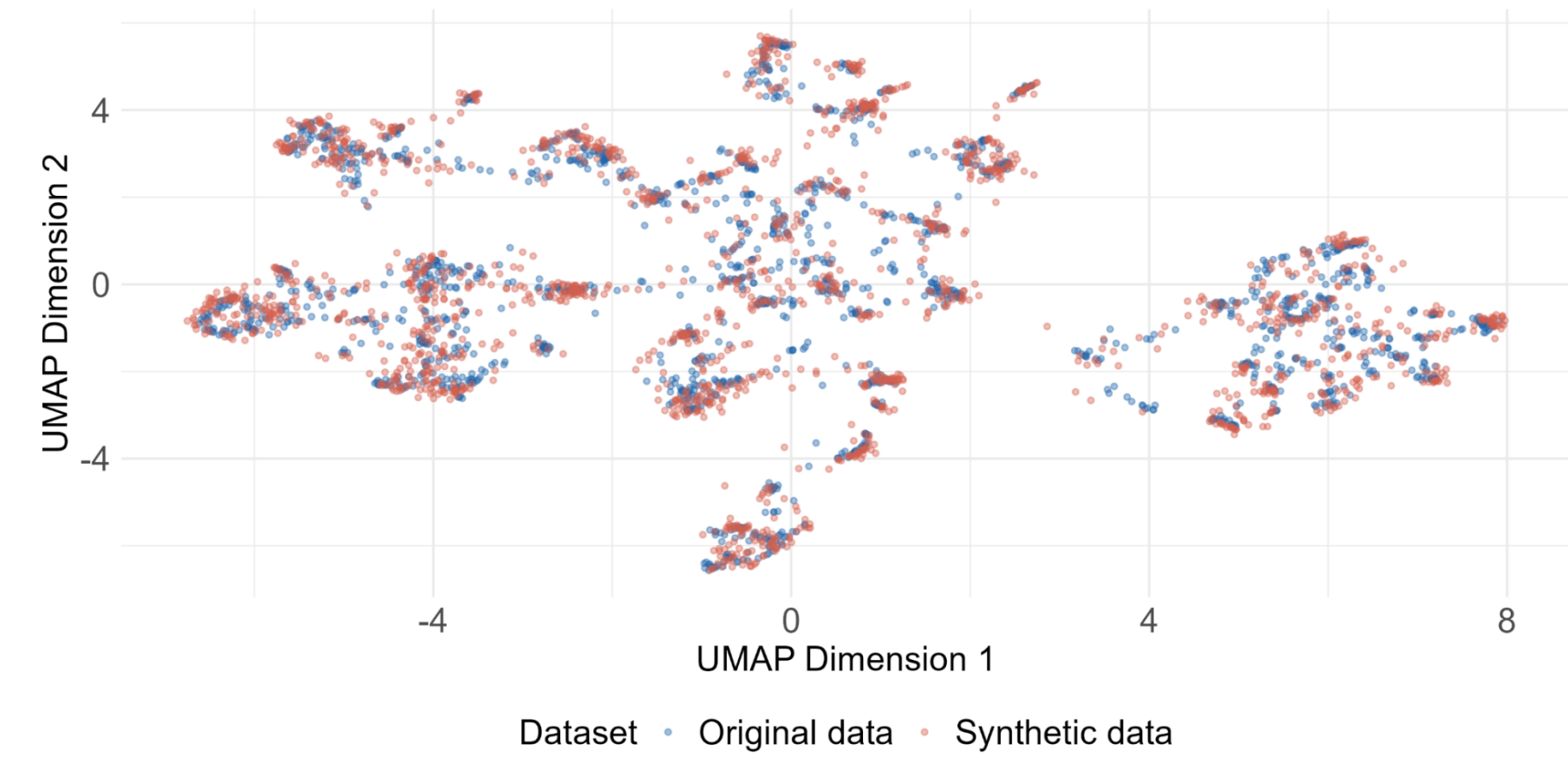


Figure 5: UMAP plot for METABRIC dataset (average seed, UMAP seed 10).

- For all datasets, low average mean ΔCC values indicated good preservation of pairwise correlations between continuous variables across multiple synthetic replicates (Table 3). From the Table 3 it follows that average ΔCC improves with increasing sample size across standard datasets.
- For the PBC dataset, mean ΔCC = 0.038 (95% CI: 0.036–0.039) and maximal ΔCC = 0.100 (95% CI: 0.095–0.110), indicating adequate preservation of multivariate correlation structure.
- PBC sits at the lower boundary of the RF method's applicability range. The minimum recommended sample size is $100 + 10 \times (\text{number of variables})$, yielding a threshold of 230 rows for PBC - which likely accounts for its comparatively elevated ΔCC values.

Conclusions

- Presented virtual population generation tool includes data synthesis and detail diagnostics.
- RF parameters (tree depth, node size, number of trees) were tuned to preserve univariate and multivariate distributions in clinical datasets.
- Inter-variable correlations were preserved by optimizing the variable generation order and selecting synthetic replicates against a target ΔCC threshold (default value: 0.1).
- Exact duplication - a known artifact of RF-based synthesis - was resolved via post-processing noise addition.
- The developed tool - *sg_vpop_est* function of **SimuRg** R package - is robust and useful for synthetic data generation in quantitative systems pharmacology including the design and execution of virtual clinical trials.

References

- Pasculli G et al. (2025). CPT Pharmacometrics Syst Pharmacol, 14(5), 840–852.
- Tibshirani RJ and Efron B (1993). Monographs on Statistics and Applied Probability, 57(1), 1–436.
- Fayette L, Brendel K and Mentré F (2026). Comput Stat Data Anal, 221, 108379.
- Zwep LB, Guo T, Nagler T et al. (2024). Clin Pharmacol Ther, 115(4), 795–804.
- Miletic M and Sariyar M (2024). Appl Sci, 14(14), 5975.
- Hundepool A et al. (2010). ESSnet Statistical Disclosure Control Handbook, 2023–12.
- Murtaza H, Ahmed M, Khan NF et al. (2023). Comput Sci Rev, 48, 100546.
- Pezoulas VC et al. (2024). Comput Struct Biotechnol J, 23, 2892–2910.
- Nowok B, Raab GM and Dibben C (2016). J Stat Softw, 74(11).
- Hammer SM et al. (1997). N Engl J Med, 337(11), 725–733.
- Schumacher M et al. (1994). J Clin Oncol, 12(10), 2086–2093.
- Pereira B et al. (2016). Nat Commun, 7, 11479.
- Dickson RE, Grambsch PM, Fleming TR et al. (1989). Hepatology, 10(1), 1–7.
- Massey FJ (1951). J Am Stat Assoc, 46(253), 68–78.
- Apellániz PA, Jiménez A, Galende BA et al. (2024). arXiv, 2405.07822v2.
- McInnes L, Healy J and Melville J (2020). arXiv, 1802.03426.

Poster III-087, 34th PAGE meeting
Dubrovnik, Croatia

Alina Melnikova
alina.melnikova@msdecisions.tech

