DOI: https://doi.org/10.15276/aait.09.2026.19
UDC 004.8

Synthetic data generation from small real-world datasets for biological age estimation

Volodymyr H. Slipchenko[1)]
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com. Scopus Author ID 59212166600
Liubov H. Poliahushko[1)]
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Volodymyr I. Rudyk[1)]
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com. Scopus Author ID 57210895061
[1)] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine

ABSTRACT

This article addresses the relevance of overcoming the shortage of high-quality clinical data required for training reliable machine learning models for biological age estimation. The collection of such data is expensive, time-consuming, and restricted by privacy regulations, while small sample sizes lead to overfitting. **The aim of this research** is to compare modern synthetic data generation methods using a limited set of skeletal system biomarkers to identify the optimal approach under small-sample conditions. **The object of the study** is the methodology of generating synthetic data for small tabular biomedical records applied to estimating biological age. **The subject of the study** encompasses statistical methods and generative machine learning models used to synthesize realistic clinical information. The objectives include selecting relevant biomarkers, implementing various data generation frameworks, assessing the quality of synthetic data, and evaluating privacy risks. The study evaluates the bootstrap resampling method, the synthetic minority oversampling technique, conditional tabular generative adversarial networks, tabular variational autoencoders, and the Gaussian copula approach. Validation relies on nonparametric statistical tests for distributional similarity, geometric proximity measurements, nearest neighbor privacy evaluations, and cross-validation testing. **The results** demonstrate that basic resampling fails to protect data privacy, while linear oversampling severely distorts chronological age distributions and lacks sample uniqueness. Deep generative adversarial networks exhibit mode collapse under small-sample conditions, shown by significant error discrepancies between training and testing scenarios. In contrast, the statistical copula approach provides the highest fidelity in maintaining complex multivariate correlations. Additionally, probabilistic autoencoders prove exceptionally capable of reproducing marginal distributions. Both of these successful methods effectively maintain privacy. **The scientific novelty** lies in the development of a comprehensive comparative framework for evaluating generative models under strict small-data constraints. **The practical value of the study** consists in providing clear recommendations for researchers: the Gaussian copula approach should be prioritized when preserving exact physiological relationships between features is critical, whereas tabular variational autoencoders are optimal when accurate population-level distributions are strictly required.

Keywords: Synthetic data; biological age; generative models; machine learning; skeletal biomarkers; data privacy

***For citation*: Slipchenko V. H., Poliahushko L. H., Rudyk V. I. "Synthetic data generation from small real-world datasets for biological age estimation". *Applied Aspects of Information Technology* 2026; Vol.9 No.3: 283–295. DOI: https://doi.org/10.15276/aait.09.2026.19**

INTRODUCTION

The rapid integration of machine learning (ML) into personalized medicine marks a significant change in modern gerontology. It provides new opportunities for predicting health status. Biological Age (BA) estimation has become a key measure because chronological age often does not accurately reflect a person's true functional state [1], [2]. In this context, assessing the skeletal system characterized by Bone Mineral Density (BMD), Trabecular Bone Score (TBS), and Fracture Risk (FRAX), is a fundamental biomarker of systemic aging [3]. Accurately modeling these parameters is essential for early detection of osteoporosis and developing preventive anti-aging strategies.

However, putting these predictive models into practice is greatly limited by the amount and quality of training data available. In biomedical research, collecting high-quality clinical data is expensive and time-consuming. This process is strictly restricted by ethical regulations and patient confidentiality. As a result, many studies rely on small datasets [4], [5], which increases the risk of overfitting and reduces the generalizability of the models [6]. This remains one of the main obstacles to applying ML methods in real clinical settings.

ANALYSIS OF EXISTING RESEARCH AND PUBLICATIONS

Synthetic data generation has become a strategic way to address data scarcity. Current literature outlines two main approaches to this issue.

Traditional data augmentation methods, like the Synthetic Minority Over-sampling Technique (SMOTE) [7] or Bootstrap resampling [8], are commonly used as baseline methods. While these methods are efficient, they work through linear interpolation or duplication. They often fail to capture the complex nonlinear relationships found in biological systems and do not guarantee patient privacy. On the other hand, Deep Generative Models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), offer promise in producing high-fidelity data [9].

Nonetheless, most current research focuses on evaluating these complex models using large datasets. A key unresolved issue is how stable the performance of these data-hungry models is when applied to limited clinical registries. There is insufficient systematic evidence to determine whether complex deep learning models provide a clear advantage over simpler statistical methods, like Gaussian Copula, in small-sample situations. They may also introduce statistical artifacts and privacy risks that threaten scientific integrity.

THE AIM AND OBJECTIVES OF THE RESEARCH

The object of study is the methodology of synthetic data generation for small tabular biomedical datasets. The subject of study is statistical methods and generative machine learning models as a means of synthesizing realistic clinical data for estimating biological age.

The work aims to evaluate and compare various synthetic data generation techniques to identify the most reliable approach for overcoming data scarcity in small biomedical datasets, ensuring both high data utility and patient privacy.

To achieve the goal, it is necessary to solve the following tasks:

1) select a specific set of skeletal system biomarkers for the assessment of biological age;

2) implement traditional data augmentation techniques and modern generative models (including Bootstrap, the Synthetic Minority Oversampling Technique, conditional tabular generative networks, tabular variational autoencoders, and the Gaussian Copula) on a limited clinical dataset;

3) evaluate the fidelity of the generated synthetic data by analyzing how well they preserve marginal distributions and complex multivariate correlation structures;

4) assess the privacy preservation capabilities of each applied method using nearest neighbor distance metrics;

5) analyze and compare the overall effectiveness of the methods based on cross-validation scenarios to determine the optimal approach for small-sample conditions.

MATERIALS AND RESEARCH METHODS

The experimental basis of this study was a specialized retrospective clinical dataset provided by experts from the D. F. Chebotarev Institute of Gerontology of the National Academy of Medical Sciences of Ukraine. The dataset is representative for gerontological analysis and is based on instrumental measurements of bone tissue status.

The final dataset used for modeling consists of anonymized medical records of 345 male patients aged 40 to 92 years, enabling the analysis of musculoskeletal system changes across both middle-aged and elderly populations. The mean age of the cohort is 59 years, which ensures adequate coverage of age-related skeletal degeneration patterns relevant to biological age estimation.

The dataset includes quantitative biomarkers characterizing bone mineral density, fracture risk, body composition, and trabecular microarchitecture. Detailed descriptive statistics of all variables are presented in Table 1.

To comprehensively assess the quality of synthetic data and their suitability for biological age estimation, seven complementary evaluation metrics were employed: Kolmogorov-Smirnov test with Bonferroni correction $\alpha_{corrected} = 0.05 / n$ for distributional similarity [10], [11]; Wasserstein distance for geometric proximity [12]; Pairwise Mutual Information for nonlinear dependencies [13]; Correlation Difference for linear relationship preservation [14]; TSTR (Train Synthetic, Test Real) and TRTS (Train Real, Test Synthetic) using Random Forest with 5-fold cross-validation for practical utility [15], [16]; NNDR (Nearest Neighbor Distance Ratio) for privacy assessment, with threshold > 1.0 indicating novel sample generation [17].

Five synthetic data generation methods were compared. Bootstrap randomly samples N observations with replacement [18], perfectly preserving statistical properties but generating no novel samples (zero-privacy baseline). SMOTE (Synthetic Minority Over-sampling Technique) synthesizes samples by linear interpolation between k-nearest neighbors: $x_{new} = x_i + \lambda \cdot (x_{zi} - x_i)$, where $\lambda \in [0,1]$ [19].

Table 1. **Detailed information about the dataset**

Biomarker name	Unit	Mean	SE
Body mass index (BMI)	kg/m^2	27.182	0.243
10-year probability of major osteoporotic fractures (FRAX-all)	%	3.174	0.085
10-year probability of hip fractures (FRAX-hip)	%	1.771	0.044
Lumbar spine bone mineral density (LS-BMD)	g/cm^2	1.023	0.011
Femoral right neck bone mineral density (FN-BMD-R)	g/cm^2	0.765	0.008
Femoral left neck bone mineral density (FN-BMD-L)	g/cm^2	0.767	0.008
Proximal right femur bone mineral density (Hip-BMD-R)	g/cm^2	0.955	0.009
Proximal left femur bone mineral density (Hip-BMD-L)	g/cm^2	0.964	0.009
BMD ultra-distal radius of the forearm (UDR-BMD)	g/cm^2	0.744	0.005
The trabecular bone scores (TBS)	units	1.32	0.006

***Source*: compiled by the authors**

Conditional Tabular Generative Adversarial Network (CTGAN) adapts GAN architecture for tabular data [20], employing a generator-discriminator framework trained with Wasserstein loss with gradient penalty [21]. The model uses Variational Gaussian Mixture normalization to handle multimodal biological biomarker distributions. Tabular Variational Autoencoder (TVAE) models data via encoder-decoder architecture [22], trained by maximizing Evidence Lower Bound $L_{TVAE} = L_{reconst} + \beta \cdot D_{KL}\left(q_\phi(z|x) \| p(z)\right)$, where $L_{reconst}$ is reconstruction loss, D_{KL} is Kullback-Leibler divergence, and β controls the trade-off between reconstruction quality and latent space structure.

Gaussian Copula is a statistical approach modeling dependency structures independently of marginal distributions [23], [24]. According to Sklar`s theorem, any multivariate distribution $F(x_1,...,x_d)$ can be represented through marginal distributions $F_i(x_i)$ and copula function $C: F(x_1,...,x_d) = C(F_1(x_1),...,F_d(x_d))$. The method transforms features to standard normal distribution, computes covariance matrix Σ, generates samples from $N(0,\Sigma)$, and transforms back to original feature space.

RESEARCH RESULTS

Initial preprocessing included cleaning, interpolation, and feature selection. The dataset was restricted to ages 40-90 (one out of-range patient removed), and records with over two missing fields were excluded. Missing values were imputed using second-order polynomial interpolation with forward and backward filling at boundaries. The final dataset size was 344 samples.

Although Pearson and Spearman analyses (Table 2) showed no significant correlation between BMI and age ($p = 0.83 > 0.1$), BMI was retained as a potential confounder [25]. This decision recognizes that body mass significantly influences BMD and fracture risk, and its inclusion is vital for the physiological validity of synthetic data. To control dataset balance, age groups were created at 10-year intervals.

Table 2. **Correlation of biomarkers with chronological age**

Biomarker name	Pearson`s correlation	Pearson`s p-value	Pearson`s t-statistics	Spearman`s correlation	Spearman`s p-value	Spearman`s t-statistics
BMI	0.012	8.30618×10^{-1}	0.2141	0.008	8.84701×10^{-1}	0.1451
FRAX-all	0.176	1.05890×10^{-3}	3.3027	0.231	1.53804×10^{-5}	4.3862
FRAX-hip	0.636	2.55892×10^{-40}	15.2229	0.726	1.17256×10^{-57}	19.5464
LS-BMD	0.129	1.68264×10^{-2}	2.4023	0.115	3.32615×10^{-2}	2.1376
FN-BMD-R	-0.192	3.45605×10^{-4}	-3.6149	-0.199	1.98422×10^{-4}	-3.7618
FN-BMD-L	-0.180	8.04475×10^{-4}	-3.3815	-0.202	1.63985×10^{-4}	-3.8112
Hip-BMD-R	-0.132	1.40984×10^{-2}	-2.4674	-0.128	1.71852×10^{-2}	-2.3944
Hip-BMD-L	-0.123	2.24059×10^{-2}	-2.2938	-0.132	1.45014×10^{-2}	-2.4571
UDR-BMD	-0.357	9.13084×10^{-12}	-7.0636	-0.354	1.35219×10^{-11}	-7.0009
TBS	-0.192	3.49754×10^{-4}	-3.6117	-0.224	2.82140×10^{-5}	-4.2448

***Source*: compiled by the authors**

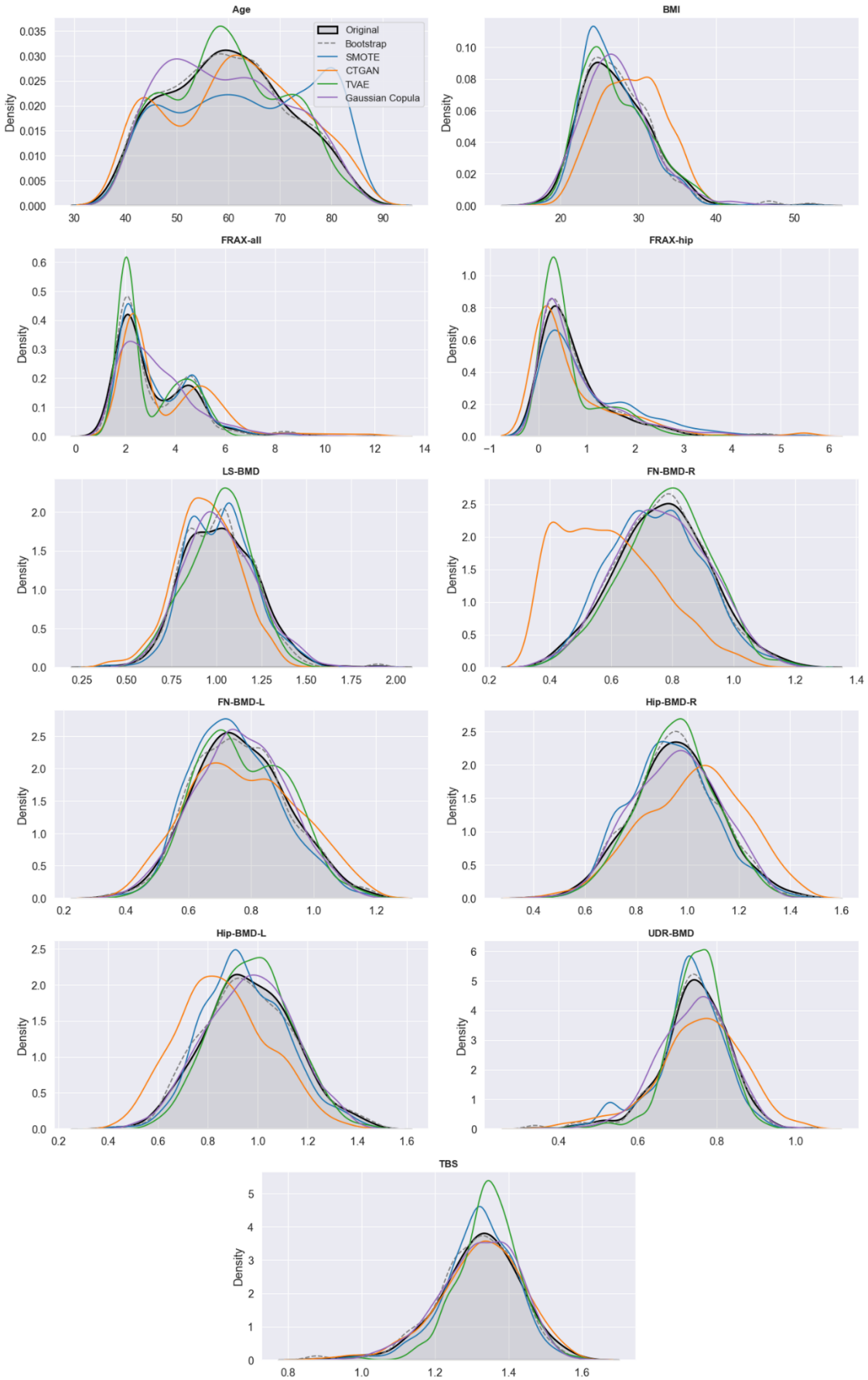


Fig. 1. **Kernel Density Estimation**
***Source*: compiled by the authors**

For the experimental comparison, five datasets were analyzed: the real dataset (N = 344) and synthetic sets (N = 1000 each) generated via Bootstrap, SMOTE, CTGAN, TVAE, and Gaussian Copula.

A key distinction was made regarding class balance: SMOTE was configured to generate 200 samples per 10-year age group, creating a perfectly balanced dataset. In contrast, the other methods aimed to reproduce the natural probability distribution and inherent class imbalances of the original clinical registry.

Prior to generation, SMOTE features were normalized to [0, 1] using MinMaxScaler to prevent scale bias; other methods utilized original scales. CTGAN and TVAE were trained for 300 epochs (batch 64), while Gaussian Copula was trained on the full dataset to capture multivariate dependencies.

Initial quality assessment was performed by visually comparing feature distributions using Kernel Density Estimation (KDE) (Fig. 1). The plots confirm that the statistical and probabilistic models closely follow the natural peaks and variances of the original clinical data, effectively capturing physiological diversity without introducing unrealistic extreme values.

For a quantitative assessment of the statistical similarity between feature distributions in the synthetic datasets and the original sample, a two-sample test was performed (Table 3). To account for multiple comparisons (11 features and 5 methods), the Bonferroni correction was applied, setting the adjusted significance threshold to $\alpha_{corrected} = 0.05/(11\times 5) \approx 0.0009$. P-values exceeding this threshold indicate no statistically significant difference between distributions.

The Bootstrap baseline demonstrated near-perfect similarity $p \approx 1$ across all features. CTGAN exhibited critical deviations $p < 0.0009$ for 7 out of 11 variables, including BMI and most bone density indicators.

Notably, SMOTE showed a significant deviation for the age variable $p \approx 0$, confirming that the linear balancing strategy distorted the natural age distribution of the cohort. Among generative models, TVAE successfully reproduced 10 out of 11 distributions (failing only on TBS), while Gaussian Copula preserved all BMD markers but deviated in FRAX-all.

Beyond statistical tests, the Wasserstein Distance (Table 4) was computed to assess geometric proximity, quantifying the minimal cost to transform synthetic distributions into real ones. TVAE achieved the lowest average distance (0.1164) among generative methods, significantly outperforming CTGAN (0.4606) and SMOTE (0.4440). Gaussian Copula showed competitive geometric fidelity (0.1347), comparable to TVAE.

The TVAE model achieved the lowest average Wasserstein distance among generative methods (0.1164), significantly outperforming CTGAN (0.4606) and SMOTE (0.4440). SMOTE's high distance for Age (3.85) further quantifies the distributional distortion introduced by its balancing procedure. Gaussian Copula showed competitive geometric fidelity (0.1347), comparable to TVAE. In practical terms, achieving a low Wasserstein distance indicates that a generative method can reliably produce synthetic patient profiles with highly realistic and clinically plausible individual characteristics, making such data safe for supplementing small medical cohorts.

Table 3. **Results of the Kolmogorov-Smirnov test (p-values) for synthetic data generation methods**

Biomarker	Bootstrap	SMOTE	CTGAN	TVAE	Gaussian Copula
Age	1.0000	0.0000	0.0779	0.8547	0.2940
BMI	0.9918	0.3181	0.0000	0.4258	0.4072
FRAX-all	1.0000	0.0092	0.0000	0.1249	0.0000
FRAX-hip	1.0000	4.5×10^{-5}	0.0000	0.0177	0.6146
LS-BMD	0.9130	0.4764	0.0000	0.0910	0.6522
FN-BMD-R	0.9998	0.1467	0.0000	0.2316	0.9946
FN-BMD-L	0.9811	0.1467	0.0705	0.4236	0.8888
Hip-BMD-R	0.9002	0.1694	0.0000	0.7258	0.7227
Hip-BMD-L	0.8892	0.1948	0.0000	0.1689	0.6979
UDR-BMD	0.9991	0.0832	0.0007	0.1321	0.0494
TBS	0.8919	0.6654	0.6208	0.0001	0.9065

***Source*: compiled by the authors**

Table 4. **Wasserstein Distance**

Biomarker	Bootstrap	SMOTE	CTGAN	TVAE	Gaussian Copula
Age	0.2105	3.8509	1.7774	0.5299	0.7651
BMI	0.2279	0.4960	2.1170	0.3101	0.3118
FRAX-all	0.0407	0.1508	0.5279	0.1944	0.2671
FRAX-hip	0.0222	0.2659	0.1678	0.1201	0.0785
LS-BMD	0.0121	0.0138	0.0746	0.0278	0.0112
FN-BMD-R	0.0046	0.0197	0.1723	0.0150	0.0066
FN-BMD-L	0.0041	0.0201	0.0254	0.0099	0.0076
Hip-BMD-R	0.0055	0.0216	0.0693	0.0125	0.0111
Hip-BMD-L	0.0091	0.0200	0.1040	0.0196	0.0087
UDR-BMD	0.0033	0.0135	0.0241	0.0132	0.0080
TBS	0.0076	0.0113	0.0073	0.0275	0.0062
Average	0.0498	0.4440	0.4606	0.1164	0.1347

***Source*: compiled by the authors**

To evaluate practical utility and privacy, TSTR and TRTS metrics were computed using a 5-fold cross-validation scheme with a Random Forest Regressor to ensure robustness and minimize bias. Additionally, the NNDR ratio was analyzed to detect potential data leakage, and Average Absolute Correlation Differences (AACD) were computed to assess the preservation of linear dependencies (Table 5).

Gaussian Copula achieved the highest correlation fidelity (AACD = 0.0290), while TVAE and Gaussian Copula both passed the privacy threshold (NNDR > 1.07), indicating the generation of novel, non-replicated samples.

The Bootstrap baseline achieved the best predictive utility (TSTR MAE ≈ 0.003) but failed the privacy check (NNDR = 0.307), confirming that samples are mere duplicates. SMOTE similarly failed the privacy threshold (NNDR = 0.853), indicating that generated samples remain too close to real records.

A significant discrepancy between TSTR and TRTS errors was observed for CTGAN (TSTR MAE 0.0022 vs. TRTS MAE 0.1193). This pattern characterizes mode collapse and overfitting: the model learns a limited set of average samples that are easy for predictive models to learn from but fail to represent the full complexity of the real distribution.

From a practical standpoint, this means that a diagnostic model trained on such collapsed synthetic data would be highly unreliable in a real-world hospital setting, as it would fail to accurately estimate the biological age of diverse patients with non-average physiological profiles.

Gaussian Copula and TVAE demonstrated the optimal trade-off, both passing the privacy check (NNDR > 1.07), indicating the generation of novel samples. Gaussian Copula excelled in preserving linear dependencies with the lowest AACD (0.0290), while TVAE provided balanced performance across all metrics. Practically,

Table 5. **Results of the quantitative evaluation of synthetic data quality**

	Bootstrap	SMOTE	CTGAN	TVAE	Gaussian Copula
TSTR_MAE	0.0031 ± 0.0027	0.0070 ± 0.0012	0.0022 ± 0.0013	0.0046 ± 0.0024	0.0053 ± 0.0068
TSTR_MSE	0.0002 ± 0.0002	0.0002 ± 0.0001	0.0001 ± 0.0001	0.0006 ± 0.0005	0.0036 ± 0.0066
TRTS_MAE	0.0482 ± 0.0066	0.1674 ± 0.0110	0.1193 ± 0.0230	0.0818 ± 0.0127	0.0816 ± 0.0154
TRTS_MSE	0.0213 ± 0.0155	0.0686 ± 0.0186	0.1064 ± 0.0667	0.0486 ± 0.0269	0.0287 ± 0.0164
NNDR	0.3070	0.8532	1.0836	1.0721	1.0737
AACD	0.0188	0.0580	0.1769	0.0701	0.0290

***Source*: compiled by the authors**

minimizing the AACD is critical in biomedical research because it ensures that well-established physiological laws such as the inverse relationship between bone mineral density and fracture risk are mathematically preserved in the synthetic dataset, preventing the generation of biologically impossible patients.

To evaluate nonlinear dependency preservation, the Pairwise Mutual Information Difference matrix was calculated (Fig. 2). The analysis of this matrix reveals that intense red zones, particularly prominent in the adversarial network's output, signify a severe disruption of the natural physiological relationships between skeletal biomarkers, which would compromise the validity of predictive risk patterns.

Greater intensity in the red color indicates a larger deviation from the original feature dependencies, with Gaussian Copula showing the highest accuracy in maintaining complex informational structures.

The CTGAN method demonstrated the poorest performance in reproducing feature dependencies (AACD = 0.1769), characterized by erratic error spikes across BMD features. In contrast, Gaussian Copula provided the most accurate reproduction of the correlation architecture, with a difference matrix showing minimal deviations (mostly within ± 0.05), confirming its suitability for small biomedical datasets.

To visualize these linear relationships, Fig. 3 presents heatmaps of the differences between the Pearson correlation matrices of real and synthetic data. The predominantly gray matrix of the copula method indicates nearly perfect preservation of clinical correlations, achieving the highest fidelity among generative models (AACD = 0.0290) with deviations mostly within ± 0.05. In sharp contrast, the deep learning approach (CTGAN) failed significantly (AACD = 0.1769). It exhibits intense red and blue clusters (deviations up to 0.59) demonstrating a failure to capture the structural integrity of bone density indicators.

TVAE (0.0701) and SMOTE (0.0580) showed moderate performance; however, SMOTE distorted age-related correlations (up to 0.12), while TVAE exhibited systemic deviations (0.10-0.22) in FRAX and BMD groups.

DISCUSSION OF THE CREATED MODELS

Our results are consistent with the findings of Kühnel et al. [26], who confirmed the high efficacy of variational autoencoders for reproducing complex medical data. However, by introducing a Bootstrap baseline, we were able to explicitly quantify the privacy-utility trade-off. While Bootstrap achieved perfect statistical fidelity (comparable to real data), its failure in the privacy metric (NNDR = 0.307) highlights the necessity of generative approaches that create novel samples rather than replicating existing ones.

The critical role of sample size becomes evident when contrasting our results with Peltier et al. [27]. While Conditional GANs were successfully applied on a dataset of 7,133 patients in their work, the same architecture in our study (N = 345) led to significant overfitting. This was empirically demonstrated by the substantial discrepancy between the low TSTR error (0.0022) and the high TRTS error (0.1193). Such a pattern suggests mode smoothing (a form of mode collapse), where the GAN learns central tendencies to minimize training loss but fails to capture the local variance and complexity required for generalization on real data.

A critical methodological finding concerns SMOTE. Although widely recommended for dataset balancing, our rigorous validation with Bonferroni correction revealed that SMOTE significantly distorted the Age distribution (p < 0.001). Furthermore, its NNDR of 0.85 indicates that the generated samples remain geometrically too close to the original records, posing privacy risks often overlooked in standard clinical ML pipelines.

In small data regime, Gaussian Copula and TVAE emerged as superior alternatives. Gaussian Copula proved to be the most robust method for preserving the correlation structure (average absolute correlation difference = 0.029), effectively modeling the physiological dependencies between biomarkers. Meanwhile, TVAE demonstrated the highest geometric fidelity, making it ideal for reproducing marginal distributions. Importantly, both methods passed the privacy threshold (NNDR > 1.07), generating novel data points that respect the underlying manifold without mere replication.

These results empirically confirm the warnings expressed by Pop et al. [28] regarding the risks of uncontrolled use of generative AI. The emergence of statistical artifacts and phantom correlations in CTGAN illustrates a threat to scientific integrity. Our study demonstrates that for limited clinical datasets, transparent statistical methods or stable probabilistic models provide a safer and more scientifically valid alternative to deep-learning black-box models.

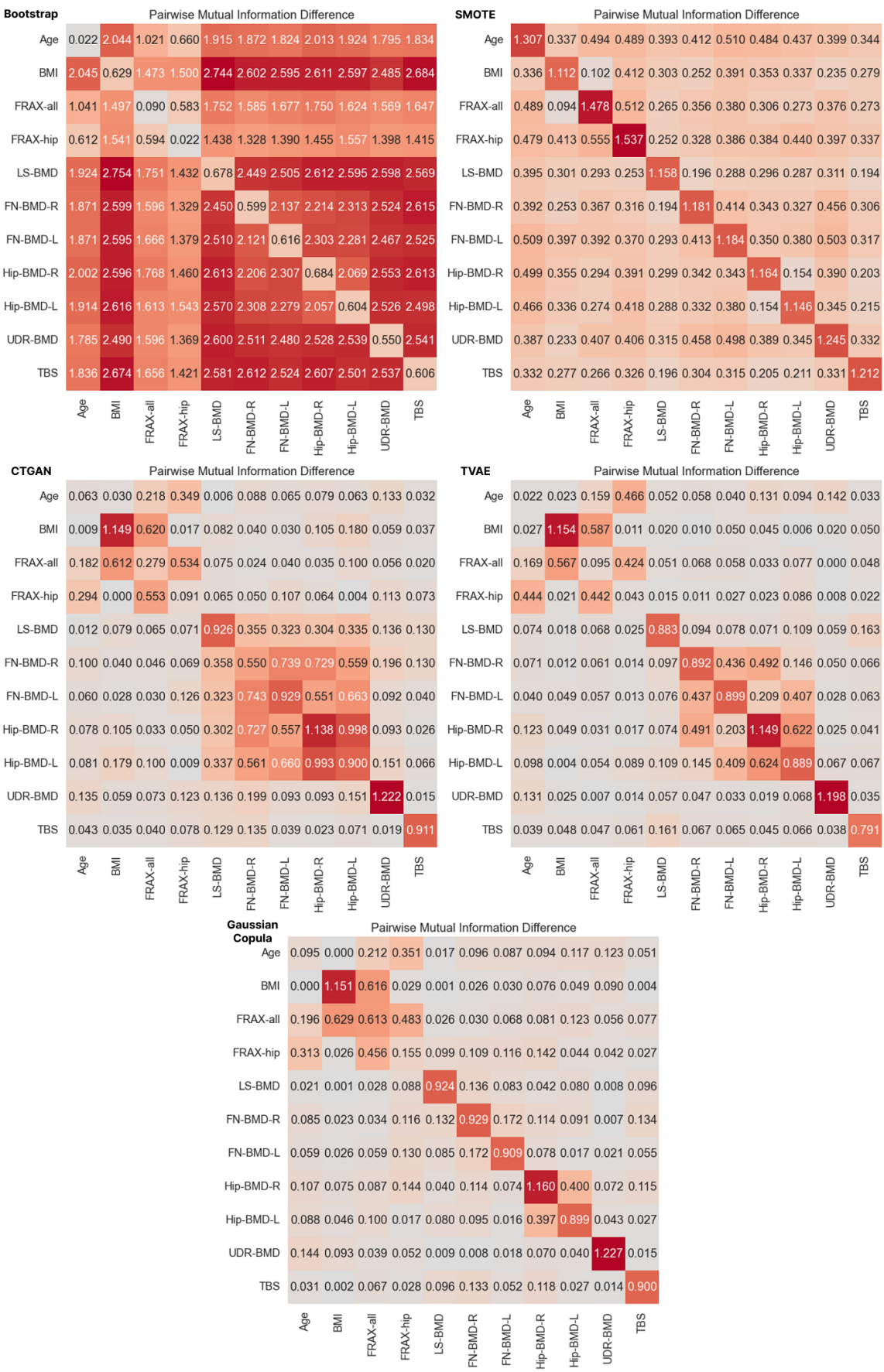


Fig. 2. **Pairwise Mutual Information Difference**
***Source*: compiled by the authors**

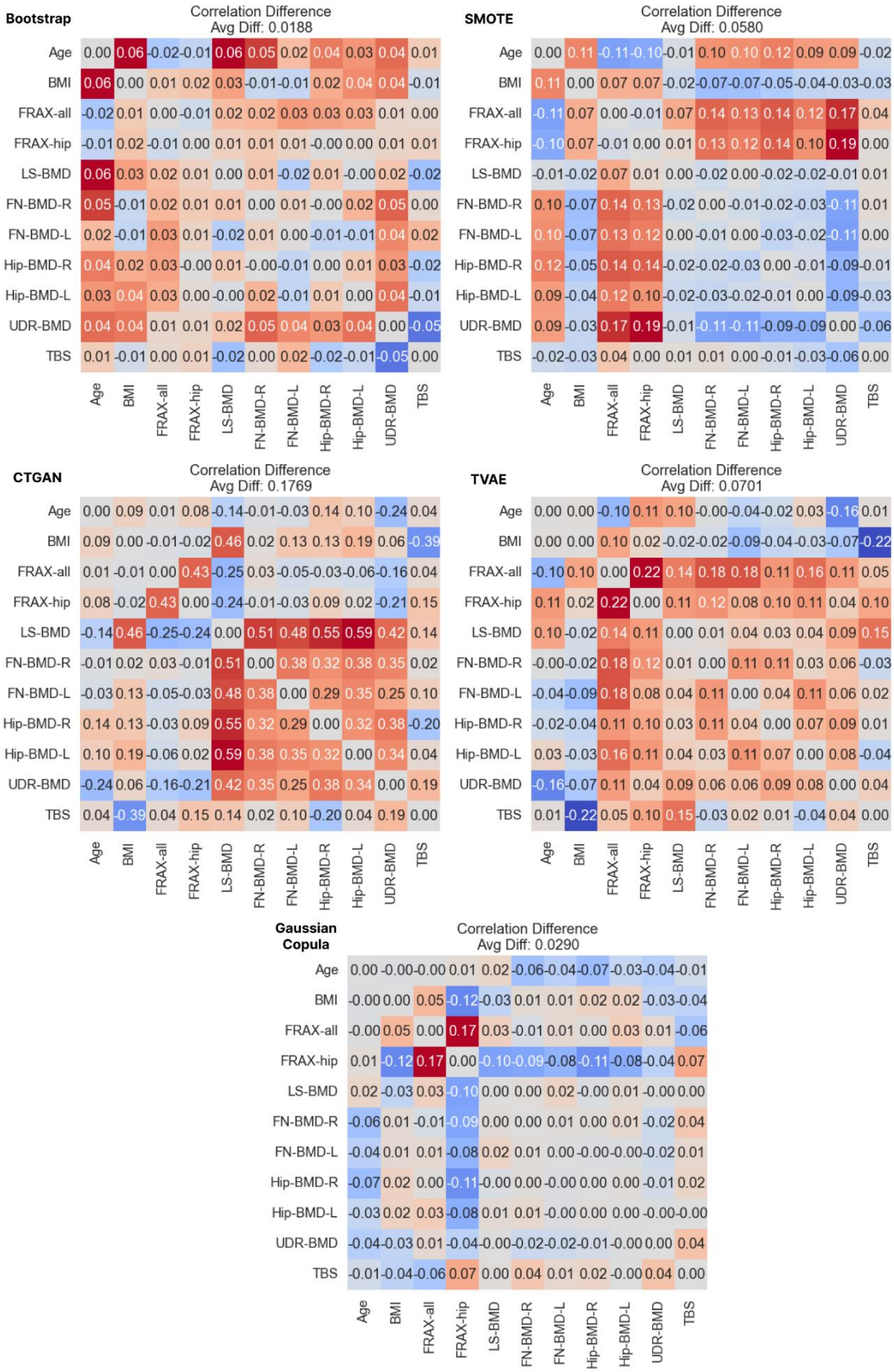


Fig. 3. **Correlation Difference Heatmap**

***Source*: compiled by the authors**

Our research has objective limitations. First, the sample size was restricted to 345 participants, which, while sufficient for the specific aim of investigating small data scenarios, limits the broader generalizability of the trained ML models. Second, the cohort consisted exclusively of male patients. Given that osteoporosis prevalence and bone aging patterns differ significantly by sex, the optimal generative parameters for female cohorts may vary. Third, the study focused on tabular data and did not evaluate the synthesis of medical imaging (DXA scans), which requires different architectural approaches.

The conditions of martial law in Ukraine directly influenced the temporal scope of this study. Data collection, which commenced in January 2017, was prematurely terminated in January 2022, as the onset of the full-scale invasion made further patient recruitment impossible. Consequently, this dataset serves as a unique pre-war baseline for biological age estimation in the studied population. It provides a reference standard that, in the presence of analogous data collected during the martial law period, will enable future comparative research to quantify the potential acceleration of aging processes induced by chronic war-related stress.

Also, future research will focus on expanding the validation framework to include female cohorts to assess gender-specific differences in generative modeling. A separate direction will involve the external validation of biological age estimation models trained on synthetic data using independent clinical cohorts from other medical centers.

CONCLUSIONS

The comparative analysis of synthetic data generation techniques on a limited biomedical dataset successfully identified the most reliable approaches for overcoming data scarcity in biological age estimation.

1. Deep learning adversarial networks and traditional linear oversampling methods proved inadequate for small clinical registries, suffering from severe mode collapse, age distribution distortion, and poor privacy preservation. Conversely, the statistical Gaussian Copula method and probabilistic tabular variational autoencoders demonstrated superior performance. They successfully generated unique, privacy-compliant synthetic samples that maintained high geometric and correlational fidelity.

2. The scientific novelty of this work consists in the systematic evaluation of the privacy-utility trade-off across varying generative architectures under strict small-data constraints. It empirically proves that complex neural architectures lose their advantage over classical statistical methods when applied to limited datasets, introducing statistical artifacts that compromise data integrity.

3. The practical relevance of the study lies in providing a justified methodology for biomedical researchers to expand limited datasets safely. Specifically, the Gaussian Copula is recommended when the preservation of precise physiological interactions between biomarkers is paramount, while tabular variational autoencoders are optimal for replicating population-level distributions without violating patient confidentiality.

Future research will focus on extending this validation framework to female cohorts to capture gender-specific bone aging patterns and evaluating the potential of diffusion models to further enhance synthetic data quality in medicine.

ACKNOWLEDGMENTS

The authors would like to thank DSc, Prof. N. Grygorieva and PhD, Sr. Research Fellow A. Musiienko from the D.F. Chebotarev Institute of Gerontology of the National Academy of Medical Sciences of Ukraine for providing the data used in this study.

REFERENCES

1. Li, S., Wen, C., Bai, X. & Yang, D. “Association between biological aging and periodontitis using NHANES 2009-2014 and Mendelian randomization”. *Scientific Reports.* 2024; 14 (1): 1–11, https://www.scopus.com/pages/publications/85192016288. DOI: https://doi.org/10.1038/s41598-024-61002-9.

2. Slipchenko, V. H., Poliahushko, L. H., Shatylo, V. V. & Rudyk, V. I. “Machine learning for human biological age estimation based on clinical blood analysis”. *Applied Aspects of Information Technologies.* 2023; 6 (4): 431–442. DOI: https://doi.org/10.15276/aait.06.2023.29.

3. Slipchenko, V., Grygorieva, N., Poliahushko, L., Musiienko, A., Rudyk, V. & Shatylo, V. “Estimation of the bone biological age using machine learning”. *EUREKA Physics and Engineering.* 2025; 1: 175–186, https://www.scopus.com/pages/publications/85217465115. DOI: https://doi.org/10.21303/2461-4262.2025.003656.

4. Chailurkit, L., Thongmung, N., Vathesatogkit, P., Sritara, P. & Ongphiphadhanakul, B. "Biological age as estimated by baseline circulating metabolites is associated with incident diabetes and mortality". *The Journal of Nutrition Health & Aging*. 2024; 28 (2): 1–6, https://www.scopus.com/pages/publications/85184704632. DOI: https://doi.org/10.1016/j.jnha.2023.100032.

5. Pisaruk, A., Shatilo, V., Grygorieva, N., Chyzhova, V., Antoniuk-Shcheglova, I., Koshel, N., Naskalova, S., Bondarenko, O., Mekhova, L., Dubetska, H., Pisaruk, L. & Shatylo, V. "Method for calculating the integrated biological age of a human". *Ageing & Longevity*. 2023; 4 (2): 45–62. DOI: https://doi.org/10.47855/jal9020-2023-2-3.

6. McCausland, T. "The bad data problem". *Research-Technology Management*. 2021; 64 (1): 68–71, https://www.scopus.com/pages/publications/85098728833.
DOI: https://doi.org/10.1080/08956308.2021.1844540.

7. Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T. & Toma, M. "SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models". *Algorithms*. 2025; 18 (1): 37, https://www.scopus.com/pages/publications/85216024285. DOI: https://doi.org/10.3390/a18010037.

8. Aloni, O., Perelman, G. & Fishbain, B. "Synthetic random environmental time series generation with similarity control, preserving original signal's statistical characteristics". *Environmental Modelling & Software*. 2024; 185: 1–37, https://www.scopus.com/pages/publications/85211127973. DOI: https://doi.org/10.1016/j.envsoft.2024.106283.

9. Ibrahim, M., Khalil, Y. A., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G. & Dumontier, M. "Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges". *Computers in Biology and Medicine*. 2025; 189: 1–42, https://www.scopus.com/pages/publications/85218947300.
DOI: https://doi.org/10.1016/j.compbiomed.2025.109834.

10. Sulewski, P. & Stoltmann, D. "Parameterized Kolmogorov-Smirnov test for normality". *Applied Sciences*. 2025; 16 (1): 1-44, https://www.scopus.com/pages/publications/105027343953. DOI: https://doi.org/10.3390/app16010366.

11. Madadizadeh, F. & Abdoli, M. "Tutorial on Bonferroni Correction as a Post Hoc analysis of a Significant Chi-Squared test: A methodological guide in food science". *Journal of Food Quality and Hazards Control*. 2025; 12: 317–324. DOI: https://doi.org/10.18502/jfqhc.12.4.20410.

12. Acciaio, B., Hou, S. & Pammer, G. "Entropic adapted Wasserstein distance on Gaussians". *Electronic Communications in Probability*. 2025; 30: 1–14, https://www.scopus.com/pages/publications/105021302357. DOI: https://doi.org/10.1214/25-ecp728.

13. Francis, D. & Sun, F. "A comparative analysis of mutual information methods for pairwise relationship detection in metagenomic data". *BMC Bioinformatics*. 2024; 25: 266, https://www.scopus.com/pages/publications/85201313004. DOI: https://doi.org/10.1186/s12859-024-05883-7.

14. Yadav, P., Gaur, M., Verma, G., Kumar, P., Fatima, N., Sarwar, S., Dwivedi, Y. R. & Madhukar, R. K. "Rigorous experimental analysis of tabular data generated using TVAE and CTGAN". *International Journal of Advanced Computer Science and Applications*. 2024; 15 (4): 1250–1262, https://www.scopus.com/pages/publications/85192094129. DOI: https://doi.org/10.14569/ijacsa.2024.01504125.

15. Pohan, H. I. "The effect of combined synthetic tabular data generated using CTGAN model with actual data on performance of DHF, Varicella, and COVID-19 recognition model". *Journal of Electrical Systems*. 2024; 20 (3): 1867–1873. DOI: https://doi.org/10.52783/jes.3797.

16. Thelagathoti, R. K., Chandel, D. S., Tom, W. A., Jiang, C., Krzyzanowski, G., Olou, A. & Fernando, M. R. "Machine learning-based ensemble feature selection and nested cross-validation for miRNA biomarker discovery in Usher syndrome". *Bioengineering*. 2025; 12 (5): 1–21, https://www.scopus.com/pages/publications/105006726649. DOI: https://doi.org/10.3390/bioengineering12050497.

17. Goyal, M. & Mahmoud, Q. H. "A systematic review of synthetic data generation techniques using generative AI". *Electronics*. 2024; 13 (17): 1–38, https://www.scopus.com/pages/publications/85203649350. DOI: https://doi.org/10.3390/electronics13173509.

18. Qin, J., Piao, J., Ning, J. & Shen, Y. "Conformal predictive intervals in survival analysis: a resampling approach". *Biometrics*. 2025; 81(2): 1–10, https://www.scopus.com/pages/publications/105006671922. DOI: https://doi.org/10.1093/biomtc/ujaf063.

19. Zhong, L. "Confronting discrimination in classification: Smote based on marginalized minorities in

the Kernel Space for imbalanced data". *arXiv*. 2024. p. 1–10. DOI: https://doi.org/10.48550/arXiv.2402.08202.

20. Zhao, Z., Kunar, A., Birke, R., Van der Scheer, H. & Chen, L. Y. "CTAB-GAN+: enhancing tabular data synthesis". *Frontiers in Big Data.* 2024; 6: 117, https://www.scopus.com/pages/publications/85182833270. DOI: https://doi.org/10.3389/fdata.2023.1296508.

21. Mtetwa, J. T., Ogudo, K. A. & Pudaruth, S. "Adaptive gradient penalty for Wasserstein GANs: Theory and applications". *Mathematics.* 2025; 13 (16): 1–28, https://www.scopus.com/pages/publications/105014466201. DOI: https://doi.org/10.3390/math13162651.

22. Carlin, L. & Benjamini, Y. "CardiCat: a Variational autoencoder for high-cardinality tabular data". *arXiv.* 2025. p. 1–15. DOI: https://doi.org/10.48550/arXiv.2501.17324.

23. Santos, D. P. & Vasconcelos, J. C. S. "Using Gaussian copulas and generative adversarial networks for generating synthetic data in beet productivity analysis". *Sugar Tech*. 2024; 27 (2): 407–417, https://www.scopus.com/pages/publications/85209687121.
DOI: https://doi.org/10.1007/s12355-024-01506-w.

24. Aich, A., Baran Aich, A. & Wade, B. "Two new generators of Archimedean copulas with their properties". *Communications in Statistics – Theory and Methods*. 2025; 54 (17): 5566–5575, https://www.scopus.com/pages/publications/85214523429.
DOI: https://doi.org/10.1080/03610926.2024.2440577.

25. Huang, Y., Liu, X., Lin, C., Chen, X., Li, Y., Huang, Y., Wang, Y. & Liu, X. "Association between the dietary index for gut microbiota and diabetes: the mediating role of phenotypic age and body mass index". *Frontiers in Nutrition.* 2025; 12: 1–11, https://www.scopus.com/pages/publications/85216874608. DOI: https://doi.org/10.3389/fnut.2025.1519346.

26. Kühnel, L., Schneider, J., Perrar, I., Adams, T., Prasser, F., Nöthlings, U., Fluck, J., Moazemi, S. & Fröhlich, H. "Synthetic data generation for a longitudinal cohort study – evaluation, method extension and reproduction of published data analysis results". *Scientific Reports.* 2024; 14 (1): 1–15, https://www.scopus.com/pages/publications/85196674361. DOI: https://doi.org/10.1038/s41598-024-62102-2.

27. D'Amico, S., Dall'Olio, D., Sala, C., et al. "Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology". *JCO Clinical Cancer Informatics.* 2023; 7: 1–22, https://www.scopus.com/pages/publications/85164234019. DOI: https://doi.org/10.1200/cci.23.00021.

28. Pop, M., Attwood, T. K., Blake, J. A., Bourne, P. E., Conesa, A., Gaasterland, T., Hunter, L., Kingsford, C., Kohlbacher, O., Lengauer, T., Markel, S., Moreau, Y., Noble, W. S., Orengo, C., Ouellette, B. F. F., Parida, L., Przulj, N., Przytycka, T. M., Ranganathan, S. & Schwartz, R. "Biological databases in the age of generative artificial intelligence". *Bioinformatics Advances*. 2024; 5 (1): 1–7, https://www.scopus.com/pages/publications/105001937636. DOI: https://doi.org/10.1093/bioadv/vbaf044.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 18.04.2026
Received after revision 10.06.2026
Accepted 17.06.2026

DOI: https://doi.org/10.15276/aait.09.2026.19
УДК 004.8

Генерація синтетичних даних на основі малих реальних наборів даних для оцінювання біологічного віку

Сліпченко Володимир Георгійович [1)]
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com. Scopus Author ID 59212166600
Полягушко Любов Григорівна[1)]
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Рудик Володимир Іванович[1)]
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com. Scopus Author ID 57210895061
[1)] Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна

АНОТАЦІЯ

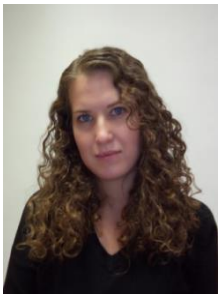
У статті розглянуто проблему подолання нестачі високоякісних клінічних даних, необхідних для навчання надійних моделей машинного навчання для оцінювання біологічного віку. Збір таких даних потрібує значних ресурсів і обмежується вимогами конфіденційності, а малі обсяги вибірок призводять до перенавчання моделей. **Метою дослідження** є порівняння сучасних методів генерації синтетичних даних на обмеженому наборі біомаркерів кісткової системи для визначення оптимального підходу в умовах малих вибірок. **Об'єктом дослідження** є методологія генерації синтетичних даних для малих табличних біомедичних наборів даних у задачах оцінювання біологічного віку. **Предметом дослідження** є статистичні методи та генеративні підходи машинного навчання, що застосовуються для синтезу реалістичної клінічної інформації. Завдання дослідження включають вибір релевантних біомаркерів, програмну реалізацію різних алгоритмів генерації, оцінку якості синтетичних даних та аналіз ризиків порушення приватності. У дослідженні оцінюються метод Bootstrap, метод синтетичного перенасичення вибірки, умовні табличні генеративно-змагальні мережі, табличні варіаційні автоенкодери та підхід на основі гаусової копули. Валідація спирається на непараметричні статистичні тести для оцінки подібності розподілів, вимірювання геометричної близькості, оцінювання приватності на основі найближчого сусіда та перехресне тестування. **У результаті** дослідження встановлено, що базовий метод Bootstrap не забезпечує належного рівня приватності, тоді як метод синтетичного перенасичення вибірки суттєво спотворює розподіл хронологічного віку та не забезпечує унікальності згенерованих зразків. Генеративно-змагальні мережі в умовах малих вибірок стикаються з колапсом мод, що підтверджується значною розбіжністю помилок між протилежними сценаріями оцінювання. Натомість підхід на основі гаусової копули забезпечує найвищу точність збереження складних кореляційних залежностей між ознаками, а табличні варіаційні автоенкодери ефективно відтворюють маргінальні розподіли. Обидва підходи успішно проходять перевірки на збереження приватності. **Наукова новизна** полягає у формуванні комплексної порівняльної методології оцінювання генеративних моделей в умовах малих вибірок. **Практичне значення** отриманих результатів полягає у формуванні рекомендацій щодо вибору методів: використання гаусової копули є доцільним у випадках, коли критично важливим є збереження взаємозв'язків між ознаками, а табличні варіаційні автоенкодери доцільно застосовувати для точного відтворення розподілів на рівні популяції.

Ключові слова: синтетичні дані; біологічний вік; генеративні моделі; машинне навчання; кісткові біомаркери; приватність даних

ABOUT THE AUTHORS

Volodymyr H. Slipchenko - Doctor of Engineering Sciences, Professor, Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0000-0002-3405-0781; ddpolytechnic2016@gmail.com. Scopus Author ID 59212166600
Research field: Development of integrated monitoring systems for ecological, energy, and economic factors; modeling for medical and biological systems; development of automated hardware and software systems; machine learning; neural networks

Сліпченко Володимир Георгійович - доктор технічних наук, професор кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна



Liubov H. Poliahushko - Candidate of Engineering Sciences, Associate Professor, Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0000-0003-3287-8523; liubovpoliagushko@gmail.com. Scopus Author ID 58246840200
Research field: Modeling for medical and biological systems; development and implementation of automated hardware and software systems; machine learning; neural networks

Полягушко Любов Григорівна - кандидат технічних наук, доцент кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна

Volodymyr I. Rudyk - PhD Student, Department of Digital Technologies in Energy. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0009-0004-4774-6579; rudykviv@gmail.com. Scopus Author ID 57210895061
Research field: Biological age; machine learning; neural networks

Рудик Володимир Іванович - аспірант кафедри Цифрових технологій в енергетиці. Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Берестейський, 37. Київ, 03056, Україна