

From Marginals to Microdata: A cGAN-based Projection for Population Synthesis

Konstantin Novikov^{*1}, Nina Gorodnova^{†2}, Alexei Romanyukha^{‡1}

¹*Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow, Russia*

²*MerlionTech LLC, Krasnogorsk, Moscow, Russia*

Abstract. We present a novel framework for synthetic population generation using conditional Generative Adversarial Networks (cGANs) to infer individual-level health microdata from population-level marginals. Specifically, our approach synthesizes realistic microdata for target populations where only aggregate statistics are available, such as age-sex distributions and disease prevalence. To address the challenges of data scarcity and privacy constraints, we develop a cGAN architecture that captures and transfers complex intervariable relationships (e.g., age-disease and disease-disease dependencies) from a source population with microdata to a target population described by marginals. A hybrid loss function enforces the fidelity to the target marginals while preserving the epidemiological realism of the generated samples. We evaluated our method using the 2023 Behavioral Risk Factor Surveillance System (BRFSS) dataset from the USA, demonstrating strong alignment with real-world distributions across multiple states. The model accurately replicates disease co-occurrence patterns and age-disease correlations, even though these were not part of the conditioning data. Our results suggest that this method can enable scalable and privacy-preserving synthetic data generation, with promising applications in public health modeling and agent-based simulation, particularly in regions lacking detailed individual-level data.

Key words: *synthetic population, deep learning, cGAN, microdata, individual-level database*

INTRODUCTION

In population studies, data may be available at different levels of detail. In most cases, aggregated (marginal) statistics describe overall distributions in a population, such as age–sex structure (e.g., from census data) or disease prevalence, but do not capture how these characteristics are combined within individuals. In contrast, microdata represent individual-level records, where multiple attributes are observed jointly for each person, although such data are less commonly available. The distinction between these two representations is crucial for population synthesis, as generating realistic microdata from marginals requires reconstructing the underlying dependencies between variables.

This challenge is closely related to the broader problem of synthetic data generation, which has become increasingly important across diverse domains, including healthcare [1], finance [2], climate science [3], and transportation [4]. While real-world microdata is ideal for analysis, it is often constrained by issues such as limited availability, high acquisition costs, and stringent privacy regulations. As a result, the development of realistic synthetic datasets has emerged as a practical solution to support modeling, decision-making, and simulation tasks.

^{*}konst.novikov@gmail.com

[†]gorodnova.mipt@gmail.com

[‡]eburg101@mail.ru

In public health, synthetic populations are foundational for individual-level simulations, such as agent-based models (ABMs) of disease transmission [5, 6]. These models rely on detailed microdata to simulate health outcomes and intervention strategies. Since advanced age and the presence of chronic conditions are key risk factors for many diseases [7], it is essential that synthetic datasets accurately reflect not only basic demographics but also the multimorbidity profile of individuals—that is, the presence of two or more coexisting chronic diseases [8]. Importantly, age and disease status are not independent [9, 10]; correlations between different chronic conditions must be preserved to ensure epidemiological realism.

Although electronic health records (EHRs) could serve as a basis for such microdata, their use is typically restricted due to privacy concerns and may not provide a comprehensive or representative view of the target population. In some regions, such data may be entirely unavailable. Classical population synthesis methods, such as synthetic reconstruction and combinatorial optimization [11, 12], require either detailed microdata or accurate joint distributions of attributes, both of which may be inaccessible. These techniques rely on iterative procedures (e.g., Iterative Proportional Fitting) or on duplicating individuals from existing samples to align with target marginals [13, 14, 12].

More advanced approaches—such as those based on Markov models, Bayesian networks, and deep generative models—have been developed to capture attribute dependencies. Recently, deep learning-based methods, including variational autoencoders (VAEs) [15], generative adversarial networks (GANs) [16], and transformer-based models [17], have shown promise in this context. However, nearly all existing methods assume access to at least some individual-level data from the target population.

In this work, we introduce a conditional GAN (cGAN)-based method for generating synthetic microdata using only population-level marginals. Our key assumption is that disease-disease correlations arise primarily from shared biological and pathophysiological mechanisms [18, 19], and therefore remain relatively stable across populations. This allows us to transfer such inter-variable dependencies from a source population (where microdata are available) to a target population (where only marginal statistics are known).

The proposed method synthesizes individual-level records conditioned on the age-sex distribution and age- and sex-stratified disease prevalence of the target population. Unlike traditional approaches that require full microdata, our framework leverages only marginals, making it well-suited for applications in data-scarce or privacy-restricted settings. We evaluate our model using data from the 2023 Behavioral Risk Factor Surveillance System (BRFSS), demonstrating its ability to preserve both marginal distributions and higher-order relationships such as disease co-occurrence and multimorbidity patterns.

Our contributions are threefold:

1. We design a cGAN architecture capable of synthesizing realistic health microdata based solely on population marginals.
2. We propose a hybrid loss function that balances adversarial learning with epidemiologically informed constraints to improve marginal alignment.
3. We empirically validate the model on real-world health data, showing robust performance in replicating both constrained and unconstrained statistical relationships.

This work offers a scalable and privacy-preserving approach to population synthesis, with potential applications in public health planning, disease modeling, and beyond.

1. METHODS

1.1. Data preparation

We utilized anonymized individual-level data from the 2023 Behavioral Risk Factor Surveillance System dataset—a nationally representative health survey conducted across 52 US states and territories [20]. From this dataset, 41 states were selected for model training (source populations), and 11 states were reserved for evaluation (target populations).

Each record included the following attributes:

- State of residence
- Age group (categorized into 13 bins, e.g., 18–24, 25–29, ..., 80–99 years)
- Sex (male/female)
- Binary indicators for seven chronic conditions: angina/coronary heart disease (CHD), myocardial infarction, stroke, chronic obstructive pulmonary disease (COPD), depression, kidney disease, and diabetes

To ensure consistency and data integrity, we performed the following preprocessing steps:

- Removed records with missing values
- Excluded records with non-binary sex labels to maintain compatibility with available marginal statistics
- Converted ordinal age categories to midpoint values (e.g., 18–24 $\rightarrow$ 21) for continuous modeling
- Recoded disease indicators as Boolean variables (True/False)

The final dataset encoded each individual as a 9-dimensional vector: one continuous variable (age) and eight binary variables (sex and disease statuses). State-specific subsets were serialized for reproducibility and consistency during training and evaluation.

1.2. Conditional GAN Framework for Population Synthesis

We developed a conditional Generative Adversarial Network (cGAN) to synthesize individual-level health records for target populations using only marginal statistics. The cGAN framework comprises two adversarial components:

- Generator (G): Produces synthetic microdata conditioned on population-level statistics
- Discriminator (D): Attempts to distinguish real data from synthetic samples

The generator accepts a random noise vector $\mathbf{z} \in \mathbb{R}^{50}$ and a conditioning vector $\mathbf{y} \in \mathbb{R}^{cond}$, where *cond* represents the dimensionality of marginal data (disease prevalence, age-sex distributions, and derived features). Age is modeled as a continuous output using a linear activation, while disease and sex indicators are modeled with sigmoid activations. The architecture of both generator and discriminator is illustrated in Figures 1 and 2, respectively.

To promote robustness, we added small noise ($\epsilon \sim 10^{-5}$) to real data during training, preventing the discriminator from overfitting to the type differences between continuous and binary inputs.

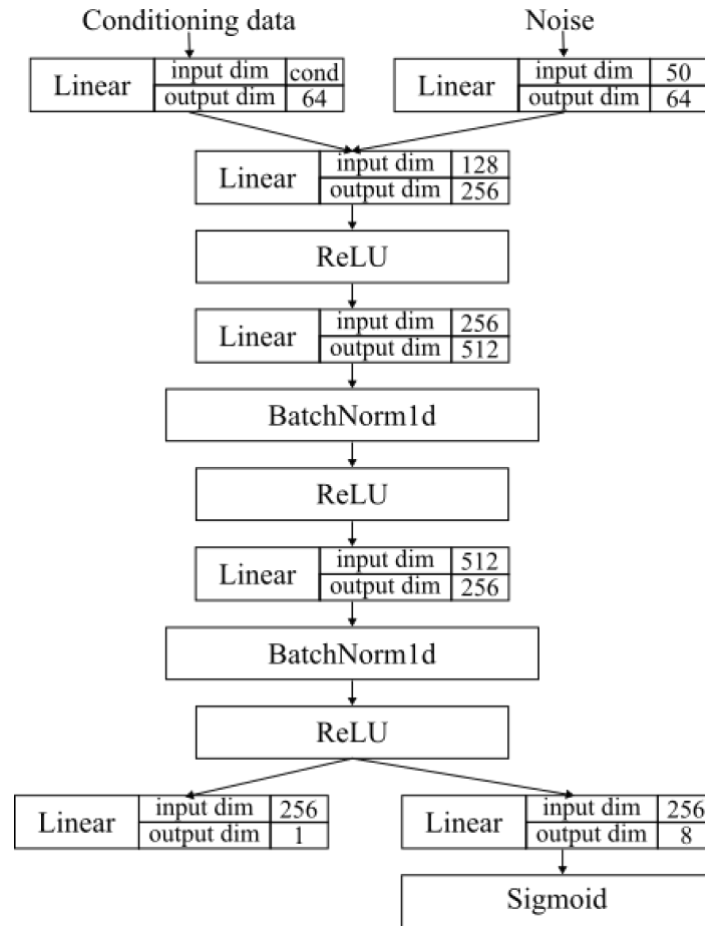


Fig. 1. Architecture of the Generator Network

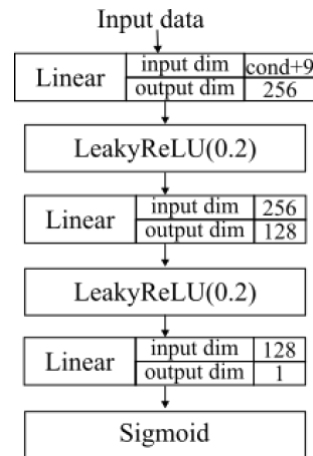


Fig. 2. Architecture of the Discriminator Network

1.3. Conditioning Mechanism

The conditioning vector $\mathbf{y}$ captures the marginal statistics of the target population and includes:

1. Age-Sex Distribution: A histogram of 13 age bins for each sex (total of 26 values)
2. Age- and Sex-Stratified Disease Prevalence: For 7 diseases across 13 age bins and 2 sex categories (total of 182 values)

To mitigate discontinuities introduced by binning, we represented age and disease marginals probabilistically. Specifically, for age, we assigned a smoothed probability s_i to each bin using a Gaussian weighting function:

$$s_i = \exp\left(-\frac{(l_i/w_i)^2}{2\sigma^2}\right)$$

where l_i is a distance to the bin center, w_i is the bin width normalized by a factor of 5 (wider for the first and last bins), and σ is a smoothing parameter. The resulting vector was normalized to ensure $\sum_i s_i$.

This probabilistic representation was applied to disease prevalence and age-sex conditioning values, enhancing the network's sensitivity to marginal transitions and improving generalization. During training, the conditioning vector $\mathbf{y}$ was held constant across each batch and concatenated with noise vector $\mathbf{z}$ to form the generator input.

1.4. Optimization

To guide training, we implemented a hybrid loss function for the generator:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{age-sex}} \mathcal{L}_{\text{age-sex}} + \lambda_{\text{prev}} \mathcal{L}_{\text{prev}},$$

where

- $\mathcal{L}_{\text{adv}}$: Adversarial loss (binary cross-entropy) from the discriminator
- $\mathcal{L}_{\text{age-sex}}$: Root Mean Square Error (RMSE) between generated and target age-sex histograms
- $\mathcal{L}_{\text{prev}}$: RMSE between generated and target age-sex-disease prevalence
- $\lambda_{\text{age-sex}} = 20$ and $\lambda_{\text{prev}} = 60$ Empirically tuned weighting factors for balancing loss function component magnitudes

We employed the Adam optimizer with learning rates of 2×10^{-4} for the generator and 10^{-6} for the discriminator. To obtain reliable marginal estimates, a large batch size of 10,000 was used during training.

Training was conducted over 140000 epochs, with a two-phase Gaussian smoothing schedule: $\sigma = 2$ for the first 70000 epochs and $\sigma = 1$ for the remainder. In each epoch, the model was trained on five randomly sampled source states from the training dataset. Evaluation was conducted on withheld states to assess generalization.

The learning process consisted of 70000 epochs with $\sigma = 2$ followed by 70000 epochs with $\sigma = 1$. Each epochs consisted of training discriminator and generator consequently on 5 randomly chosen datasets (USA states) of training set. We evaluated the model on other states, that were not used for training.

The model was implemented in Python 3.11.7 using PyTorch 2.5.1.

2. RESULTS

To evaluate the performance of our cGAN framework, we compared generated synthetic populations against real population data from 11 U.S. states (test datasets) that were excluded from training. Evaluation focused on two primary aspects:

1. Fidelity to conditioning marginals: age-sex distributions and age/sex-stratified disease prevalence
2. Preservation of unconstrained relationships: disease-disease correlations, age-disease associations, and multimorbidity profiles

2.1. Alignment with Target Marginals

2.1.1. Age-Sex Distribution

The synthetic populations closely replicated the target age-sex distributions. The Wasserstein distance between the true and generated age-sex histograms, calculated across all test datasets, was 0.0047 (95 % CI: [0.0028, 0.0062]), indicating high fidelity. For comparison, the distance between the generated population and a uniform distribution was substantially higher (0.0353, 95 % CI: [0.0300, 0.0383]), demonstrating the model's ability to accurately reproduce population structure.

Notably, in cases where real data were incomplete (e.g., certain strata missing in North Carolina), the model plausibly inferred the missing segments, demonstrating potential utility in data imputation scenarios (Fig. 3).

2.1.2. Age-Stratified Disease Prevalence

The model accurately reproduced age- and sex-stratified prevalence rates for all seven chronic diseases. As shown in Figure 4, the predicted prevalence values fell within the standard deviation bounds observed across the real test datasets in most cases. The overall RMSE across all diseases and bins was 6.89 % (95 % CI: [4.05 %, 12.68 %]), suggesting strong alignment with the conditioning inputs.

2.2. Evaluation of Unconstrained Dependencies

To assess the model's ability to preserve complex inter-variable relationships that were not part of the conditioning vector, we examined several key epidemiological patterns.

2.2.1. Disease-Disease Correlation Structure

Synthetic populations replicated pairwise disease correlations observed in the test datasets. An example correlation matrix is presented in Figure 5. The mean RMSE across all pairwise disease correlations and test states was 0.060 (95 % CI: [0.037, 0.092]). This result supports our assumption that disease co-occurrence patterns can be robustly transferred across populations using the proposed architecture.

2.2.2. Age-Disease Correlations

We also evaluated how well the model captured associations between age and individual disease presence. As depicted in Figure 6, the correlation patterns in the generated population aligned closely with those of the real datasets. The mean RMSE across all age-disease correlations was 0.038 (95 % CI: [0.022, 0.056]).

2.2.3. Multimorbidity Distribution

The distribution of individuals by the number of coexisting chronic conditions (0 to 7) was also analyzed. As shown in Figure 7, the synthetic population accurately reproduced this distribution, with an overall RMSE of 0.017 (95 % CI: [0.011, 0.027]). While the model slightly overestimated the proportion of individuals without any chronic conditions in 9 of 11 test states, the overall pattern was well preserved.

2.3. Summary

Across all evaluation dimensions—both conditioned and unconditioned—the proposed cGAN framework demonstrated strong generalization performance. It was capable of:

- accurately reconstructing population marginals;

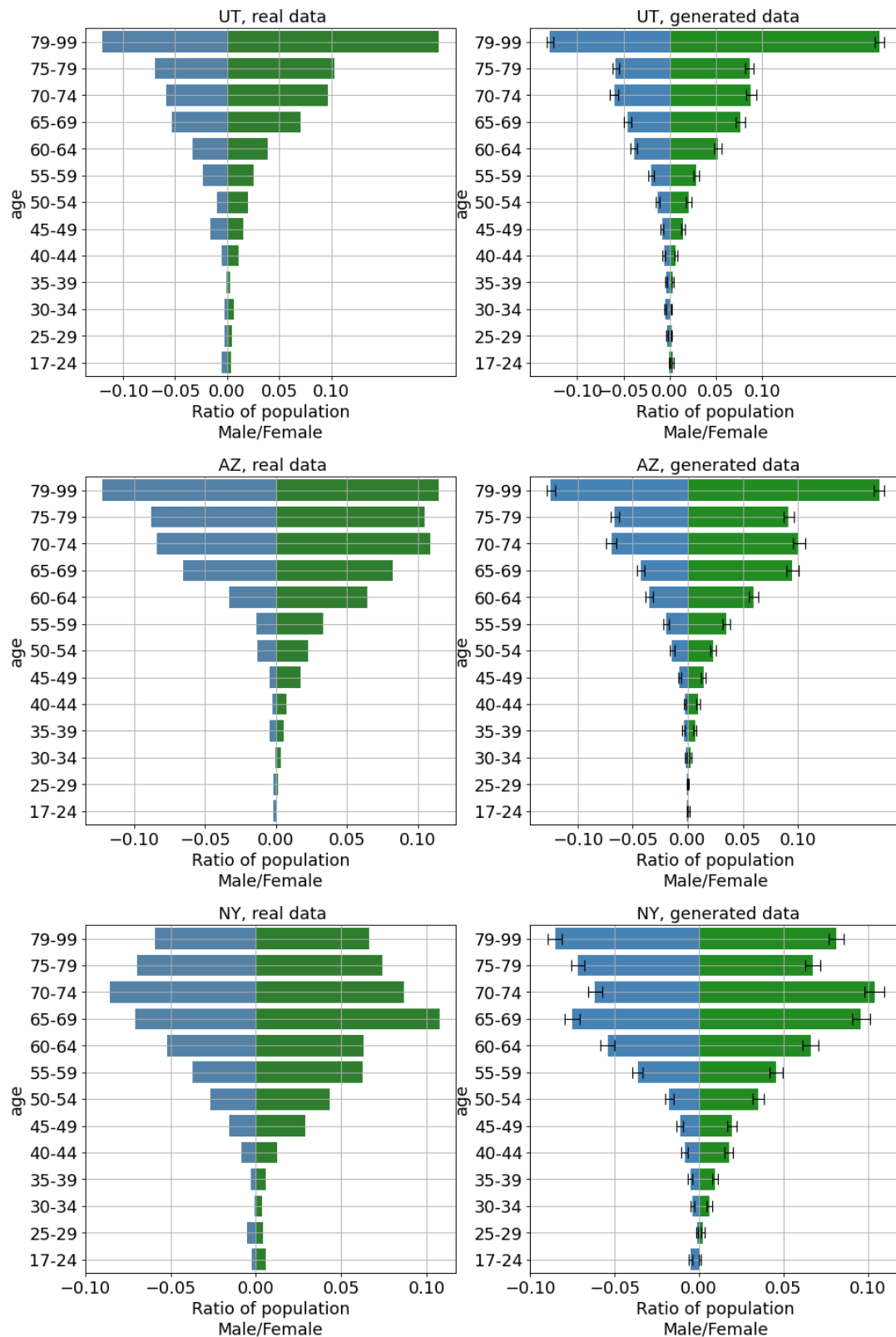


Fig. 3. Comparison of age-sex pyramids for real (left) and generated (right) populations in a sample test datasets. Whiskers represent the 95 % confidence interval based on 1,000 synthetic generations

- preserving complex dependency structures (including those not directly enforced during training);
- filling in missing or incomplete marginal data with epidemiologically plausible values.

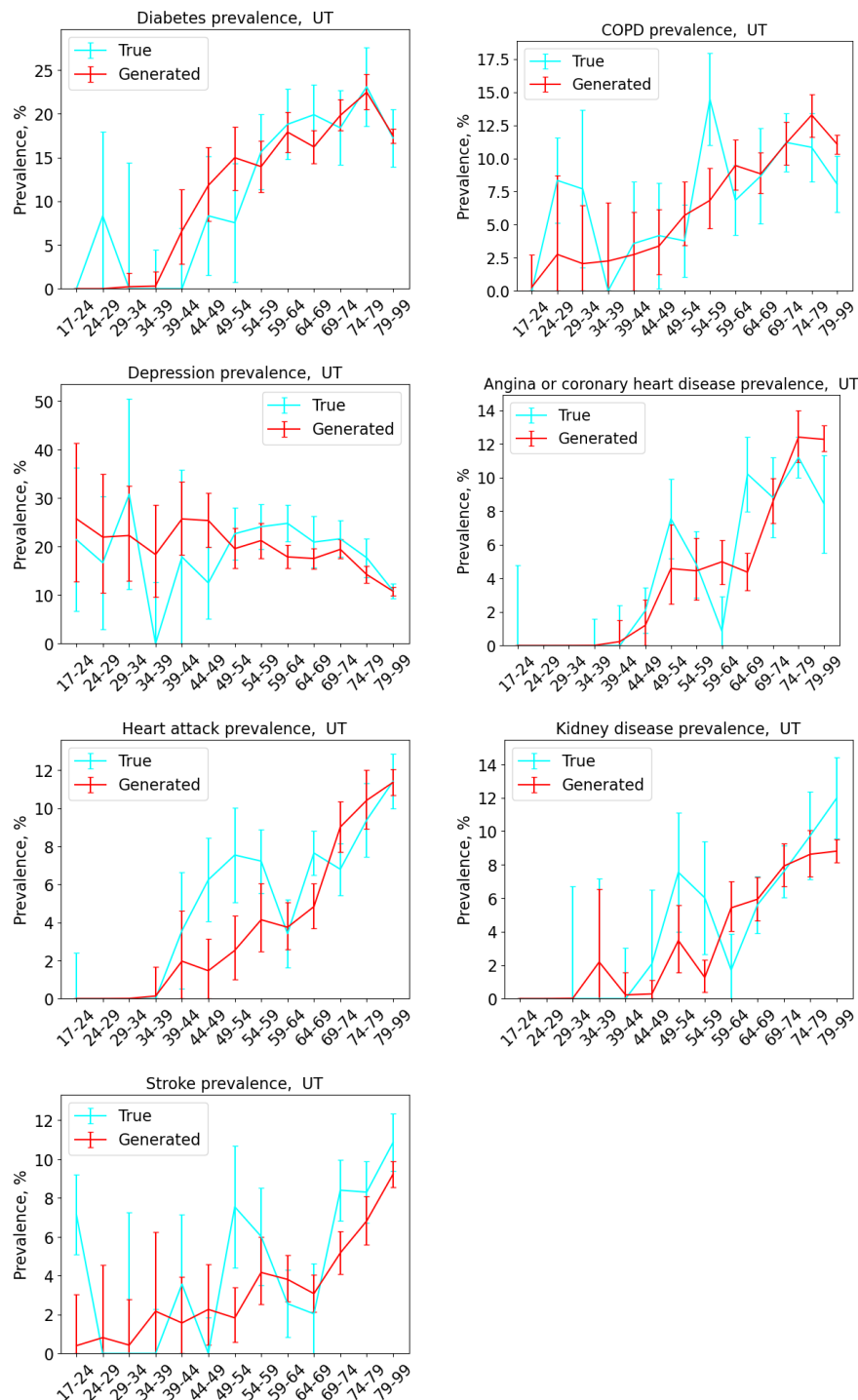


Fig. 4. Comparison of age-stratified disease prevalence between real and generated populations for an example test dataset. Whiskers on real values represent the standard deviation across all test datasets; for generated values, they represent the 95 % confidence interval over 1,000 generations

These results validate the framework's utility for generating high-fidelity synthetic populations in settings where only aggregated health statistics are available.

3. DISCUSSION

This study addresses a critical challenge in synthetic population modeling: generating realistic individual-level health microdata for a target population using only aggregated marginal statistics. By leveraging a conditional Generative Adversarial Network (cGAN)

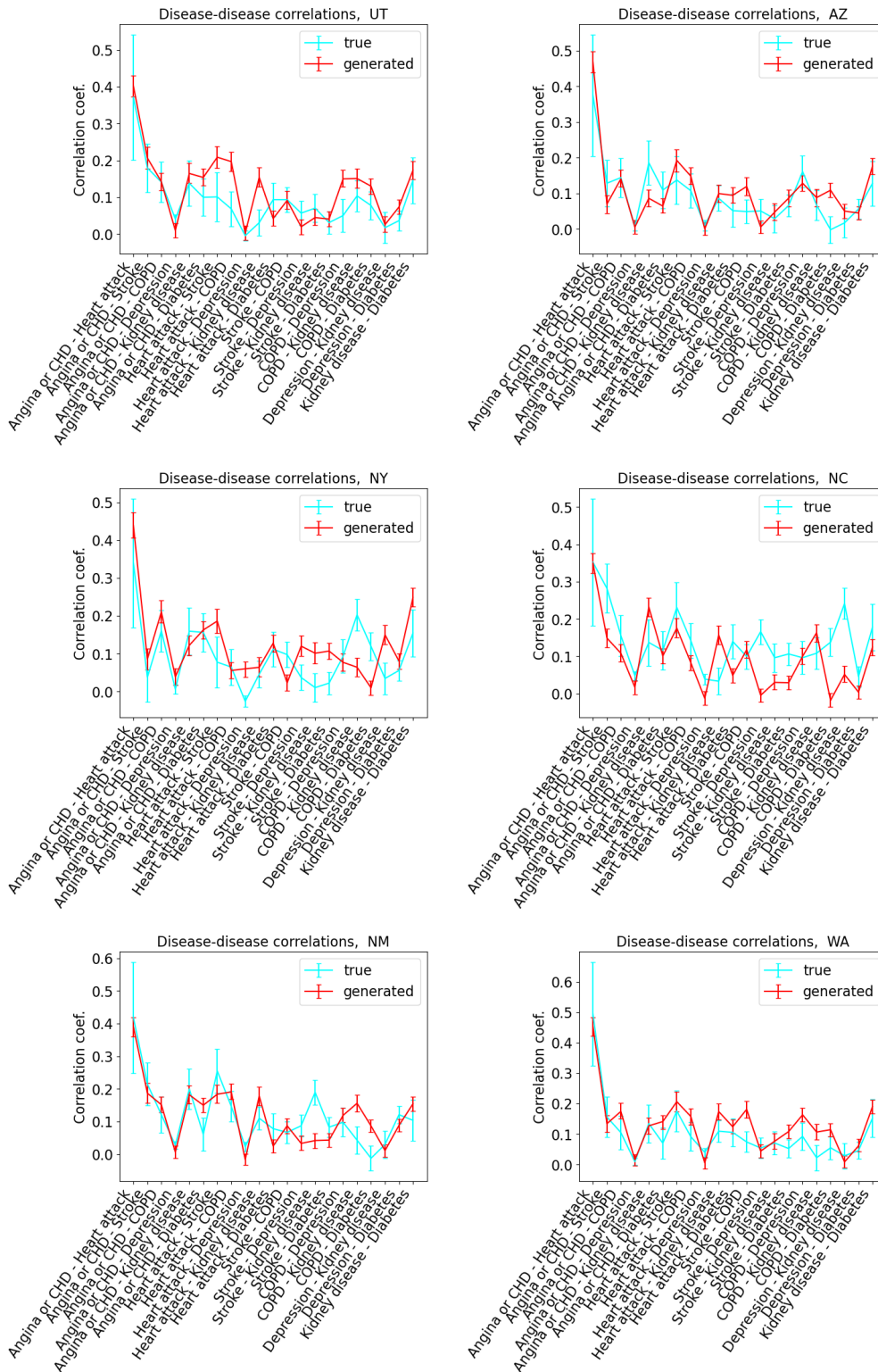


Fig. 5. Comparison of disease-disease correlation matrices for real and generated populations. For real values, whiskers indicate the standard deviation across test datasets; for generated values, they show the 95 % confidence interval estimated from 1,000 synthetic generations

architecture enhanced with epidemiological constraints, we demonstrate that key statistical and relational properties of health data—such as disease prevalence, age-disease associations, and multimorbidity patterns—can be accurately reproduced, even in the absence of full microdata.

Our framework introduces a novel capability: projecting fine-grained microdata from a

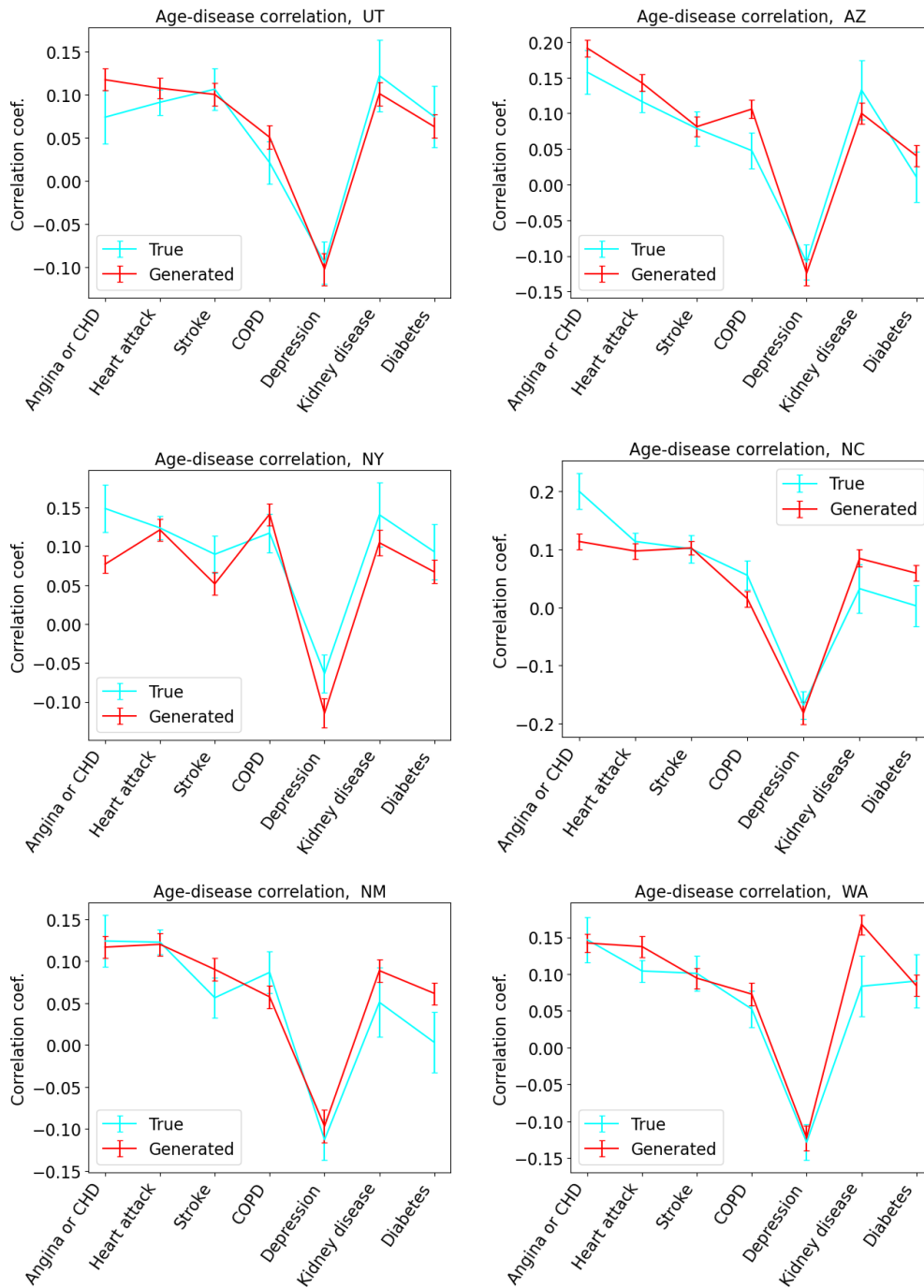


Fig. 6. Comparison of age-disease correlation coefficients between real and generated populations for example test datasets. Whiskers for real values represent standard deviation across test datasets; whiskers for generated values denote the 95 % confidence interval over 1,000 generations

source population to a target population using only age-sex pyramids and age/sex-stratified disease prevalence. Unlike traditional methods that require access to full individual-level records, our model operates under more restrictive and realistic conditions, where only summary statistics may be available due to privacy regulations or data scarcity.

The experimental results on BRFSS 2023 data show strong alignment between generated and real populations. Notably, the model not only replicates conditioned marginals with low error (e.g., prevalence RMSE $\approx 6.89\%$) but also preserves unconstrained relationships such as disease-disease correlations and multimorbidity profiles. These findings support our underlying

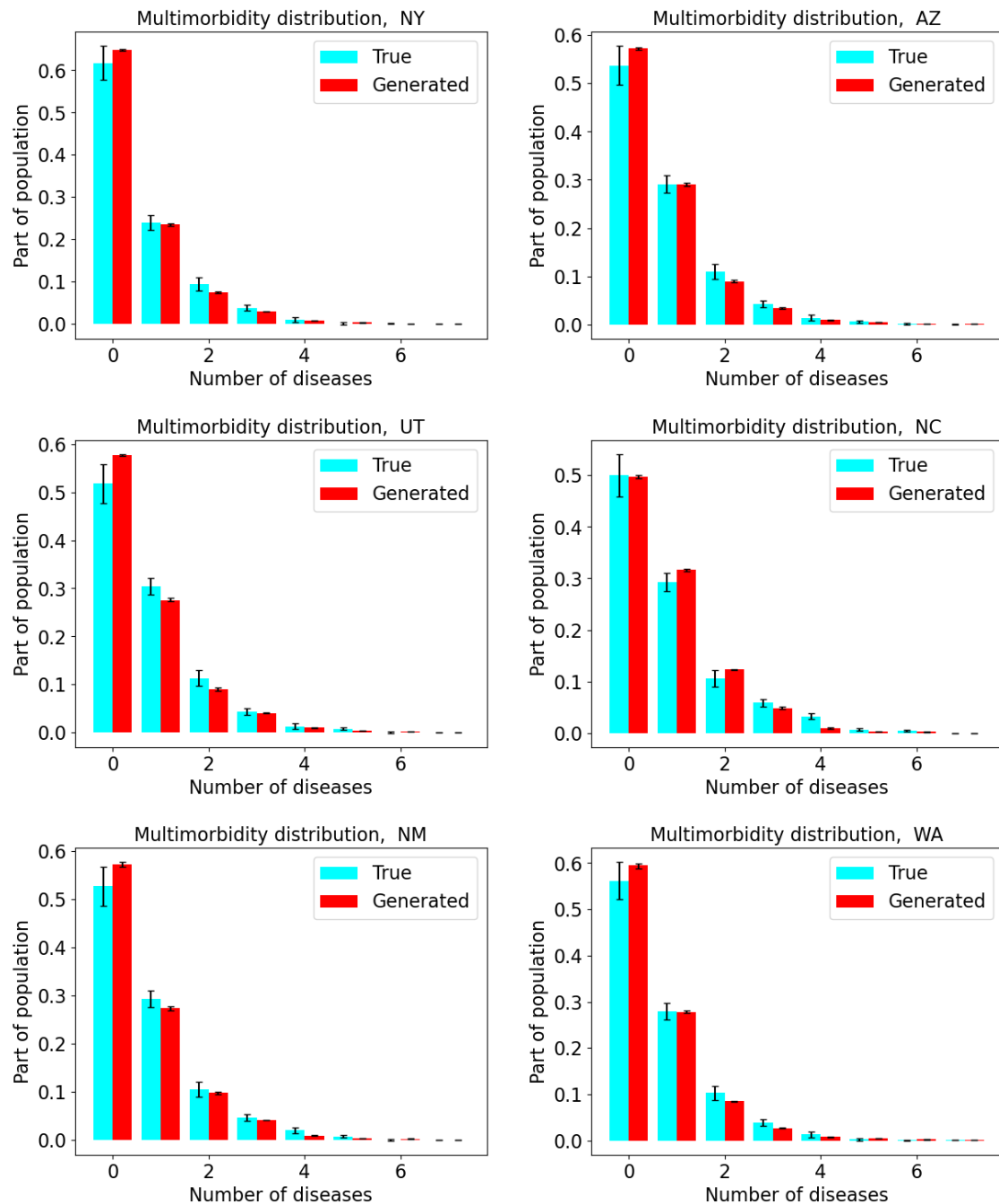


Fig. 7. Comparison of the distribution of coexisting chronic diseases per individual in real and generated populations for example test datasets. Whiskers for real values indicate the standard deviation across test datasets; whiskers for generated values reflect the 95 % confidence interval based on 1,000 generations

assumption that inter-variable relationships—particularly disease co-occurrence—are governed by stable pathophysiological mechanisms that generalize across populations.

However, certain limitations were observed. The model tends to slightly overestimate the proportion of individuals without any chronic diseases. This bias suggests that while the model captures universal patterns well, it may underrepresent subtle population-specific deviations. Addressing this issue may require incorporating additional conditioning variables (e.g., socio-economic or lifestyle factors) or dynamic adjustment mechanisms during training.

Furthermore, the method assumes that the source and target populations share similar

underlying health structures. While this assumption may hold across regions with comparable demographics and healthcare environments, it may not generalize to genetically, culturally, or environmentally distinct populations. Extending the model to explicitly account for such variation remains an important direction for future research.

Another limitation is the current lack of spatial, behavioral, or environmental covariates, which could further enhance the accuracy of regional projections. Incorporating auxiliary data sources (e.g., census income levels, geographic risk indicators) could improve the precision and relevance of the synthesized microdata for policy planning and simulation.

Despite these limitations, our approach offers several practical advantages. It is scalable, privacy-preserving, and applicable in scenarios where detailed individual-level data are unavailable. Potential applications extend beyond public health to domains such as financial modeling, urban planning, and education, where synthetic populations can be used for simulation, forecasting, and decision support without violating privacy constraints.

Finally, ethical considerations must be addressed when deploying synthetic population models. The risk of propagating biases from source datasets or misrepresenting vulnerable subpopulations calls for careful validation and governance mechanisms. Ensuring transparency in model assumptions and providing uncertainty quantification are essential for responsible use.

The authors have no conflict of interests to declare that are relevant to the content of this article. All data used in this study are publicly available from Centers for Disease Control and Prevention web page at www.cdc.gov/brfss.

REFERENCES

1. Pezoulas V.C., Zaridis D.I., Mylona E., Androutsos C., Apostolidis K., Tachos N.S., Fotiadis D.I. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*. 2024. V. 23. P. 2892–2910. doi: [10.1016/j.csbj.2024.07.005](https://doi.org/10.1016/j.csbj.2024.07.005)
2. Tkachenko N. Opportunities for synthetic data in nature and climate finance. *Frontiers in Artificial Intelligence*. 2024. V. 6. Article no. e1012439. doi: [10.3389/frai.2023.1168749](https://doi.org/10.3389/frai.2023.1168749)
3. Meyer D., Nagler T., Hogan R.J. Copula-based synthetic data augmentation for machine-learning emulators. *Geoscientific Model Development*. 2021. V. 14. No. 8. P. 5205–5215. doi: [10.5194/gmd-14-5205-2021](https://doi.org/10.5194/gmd-14-5205-2021)
4. Kapp A., Hansmeyer J., Mihaljević H. Generative models for synthetic urban mobility data: A systematic literature review. *ACM Computing Surveys*. 2023. V. 56. No. 4. Article No. 93. doi: [10.1145/3610224](https://doi.org/10.1145/3610224)
5. Kerr C.C., Stuart R.M., Mistry D., Abeyesuriya R.G., Rosenfeld K., Hart G.R., Núñez R.C., Cohen J.A., Selvaraj P., Hagedorn B. et al. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*. 2021. V. 17. No. 7. Article No. e1009149. doi: [10.1371/journal.pcbi.1009149](https://doi.org/10.1371/journal.pcbi.1009149)
6. Von Hoene E., Roess A., Kavak H., Anderson T. Synthetic population generation with public health characteristics for spatial agent-based models. *PLOS Computational Biology*. 2025. V. 21. No. 3. Article No. e1012439. doi: [10.1371/journal.pcbi.1012439](https://doi.org/10.1371/journal.pcbi.1012439)
7. Zhang J.-J., Dong X., Liu G.-H., Gao Y.-D. Risk and protective factors for covid-19 morbidity, severity, and mortality. *Clinical Reviews in Allergy & Immunology*. 2023. V. 64. No. 1. P. 90–107. doi: [10.1007/s12016-022-08921-5](https://doi.org/10.1007/s12016-022-08921-5)
8. Skou S.T., Mair F.S., Fortin M., Guthrie B., Nunes B.P., Miranda J.J., Boyd C.M., Pati S., Mtenga S., Smith S.M. Multimorbidity. *Nature Reviews Disease Primers*. 2022. V. 8. Article No. 48. doi: [10.1038/s41572-022-00376-4](https://doi.org/10.1038/s41572-022-00376-4)
9. Feng Y., Wang J., Wang Y., Helal S. Completing Missing Prevalence Rates for Multiple

- Chronic Diseases by Jointly Leveraging Both Intra- and Inter-Disease Population Health Data Correlations. In: *WWW '21: Proceedings of the Web Conference 2021*. Eds.: Leskovec J., Grobelenk M., Najork M., Tang J., Zia L. New York: Association for Computing Machinery, 2021. P. 183–193. doi: [10.1145/3442381.3449811](https://doi.org/10.1145/3442381.3449811)
10. Nesto R.W. Correlation between cardiovascular disease and diabetes mellitus: current concepts. *The American Journal of Medicine*. 2004. V. 116. No. 5. P. 11–22. doi: [10.1016/j.amjmed.2003.10.016](https://doi.org/10.1016/j.amjmed.2003.10.016)
 11. Chapuis K., Taillandier P., Drogoul A. Generation of synthetic populations in social simulations: A review of methods and practices. *Journal of Artificial Societies and Social Simulation*. 2022. V. 25. No. 2. Article No. 6. doi: [10.18564/jasss.4762](https://doi.org/10.18564/jasss.4762)
 12. Prédhumeau M., Manley E. A synthetic population for agent-based modelling in Canada. *Scientific Data*. 2023. V. 10. No. 1. Article No. 148. doi: [10.1038/s41597-023-02030-4](https://doi.org/10.1038/s41597-023-02030-4)
 13. Deming W.E., Stephan F.F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*. 1940. V. 11. No. 4. P. 427–444. doi: [10.1214/aoms/1177731829](https://doi.org/10.1214/aoms/1177731829)
 14. Lenormand M., Deffuant G. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation*. 2013. V. 16. No. 4. Article No. 12. doi: [10.18564/jasss.2319](https://doi.org/10.18564/jasss.2319)
 15. Garrido S., Borysov S.S., Pereira F.C., Rich J. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*. 2020. V. 120. Article No. 102787. doi: [10.1016/j.trc.2020.102787](https://doi.org/10.1016/j.trc.2020.102787)
 16. Choi E., Biswal S., Malin B., Duke J., Stewart W.F., Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Eds.: Doshi-Velez F., Fackler J., Kale D., Ranganath R., Wallace B., Wiens J. Boston: PMLR, 2017. Vol. 68. P. 286–305.
 17. Phattharajiranan P., Muangsin V. Generating synthetic population using transformer based networks. In: *2023 27th International Computer Science and Engineering Conference (ICSEC 2023)*. Samui Island: IEEE, 2023. P. 193–200. doi: [10.1109/ICSEC59635.2023.10329722](https://doi.org/10.1109/ICSEC59635.2023.10329722)
 18. Demela P., Pirastu N., Soskic B. Cross-disorder genetic analysis of immune diseases reveals distinct gene associations that converge on common pathways. *Nature Communications*. 2023. V. 14. Article No. 2743. doi: [10.1038/s41467-023-2743](https://doi.org/10.1038/s41467-023-2743)
 19. Li Y., Agarwal P. A pathway-based view of human diseases and disease relationships. *PLoS One*. 2009. V. 4. No. 2. Article No. e4346. doi: [10.1371/journal.pone.0004346](https://doi.org/10.1371/journal.pone.0004346)
 20. *BRFSS Survey Data and Documentation*. 2023.
URL: www.cdc.gov/brfss/annual_data/annual_2023.html (accessed: 21.04.2026).