Generative Adversarial Networks for Modeling Clinical Biomarker Profiles in Under-Represented Groups

Rahul Nair[1], Deen Mohan[1], Sandra Frank[1], Srirangaraj Setlur[1], Venugopal Govindaraju[1], and Murali Ramanathan[1]

[1]University at Buffalo

May 23, 2022

Abstract

Background: Clinical trial simulations and pharmacometric modeling of biomarker profiles for under-represented groups are challenging because the underlying studies frequently do not have sufficient participants from these groups. Objectives: To investigate generative adversarial networks (GANs), an artificial intelligence (AI) technology that enables realistic simulations of complex patterns, for modeling clinical biomarker profiles of under-represented groups. Methods: GANs consist of generator and discriminator neural networks that operate in tandem. GAN architectures were developed for modeling univariate and joint distributions of a panel of 16 diabetes-relevant biomarkers from the National Health and Nutrition Examination Survey (NHANES), which contains laboratory and clinical biomarker data from a population-based sample of individuals of all ages, racial groups, and ethnicities. Conditional GANs were used to model biomarker profiles for race/ethnicity categories. GAN performance was assessed by comparing GAN outputs to test data. Results: The biomarkers exhibited non-normal distributions and varied in their bivariate correlation patterns. Univariate distributions were modeled with generator and discriminator neural networks consisting of two dense layers with rectified linear unit-activation. The distributions of GAN-generated biomarkers were similar to the test data distributions. The joint distributions of the biomarker panel in the GAN-generated data were dispersed and overlapped with the joint distribution of the test data as assessed by three multi-dimensional projection methods. Conditional GANs satisfactorily modeled the joint distribution of the biomarker panel in the Black, Hispanic, White, and "Other" race/ethnicity categories. Conclusions: GAN are a promising AI approach for generating virtual patient data with realistic biomarker distributions for under-represented race/ethnicity groups.

Generative Adversarial Networks for Modeling Clinical Biomarker Profiles in Under-Represented Groups

Rahul Nair [1], Deen Dayal Mohan [1], Sandra Frank [2], Srirangaraj Setlur[1], Venugopal Govindaraju [1] and Murali Ramanathan [2]

[1] Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA.

[2] Department of Pharmaceutical Sciences, University at Buffalo, The State University of New York, Buffalo, NY, USA.

Corresponding Author: Murali Ramanathan, 355 Pharmacy Building, Department of Pharmaceutical Sciences, State University of New York, Buffalo, Buffalo, NY 14214-8033. (716)-645-4846 and FAX 716-829-6569. E-mail Murali@Buffalo.Edu. ORCID: 0000-0002-9943-150X.

Running Head : Generative adversarial networks for biomarkers

Keywords: Artificial intelligence, AI, generative adversarial networks, pharmacometrics.

Word Count: Title: 100 Characters, Running Head: 47 characters, Abstract: 250 words, Introduction to Discussion: 3670 words. References: 25. Tables: 1. Figures: 5.

Data availability statement: The data that support the findings of this study are openly available in NHANES at https://www.cdc.gov/nchs/nhanes/index.htm, reference number 16.

Author Contributions: Rahul Nair – Conducted experiments, data analysis, manuscript preparation. Sandra Frank – Obtained data, data analysis, manuscript preparation. Deen Dayal Mohan – Designed experiments, data analysis, manuscript preparation. Srirangaraj Setlur – Study concept and design, data analysis, manuscript preparation. Venu Govindaraju – Study oversight, manuscript review. Murali Ramanathan – Study concept and design, data analysis, manuscript preparation.

Ethics approval statement: Not applicable

Patient consent statement: Not applicable

Permission to reproduce material from other sources: Not applicable

Clinical trial registration: Not applicable

Conflict of Interest Disclosure: Rahul Nair, Sandra Frank, and Deen Dayal Mohan have no conflicts. Srirangaraj Setlur and Venu Govindaraju received unrelated research funding from the National Science Foundation, United State Postal Service, and the Intelligence Advanced Research Projects Activity agencies. received unrelated research funding from the National Science Foundation, United States Postal Service, and the Intelligence Advanced Research Projects Activity agencies. Murali Ramanathan received research funding from the National Science Foundation, Otsuka Pharmaceuticals, and the National Institutes of Health.

Funding: Support from Grant MS190096 from the Department of Defense Multiple Sclerosis Research Program for the Office of the Congressionally Directed Medical Research Programs (CDMRP) to the Ramanathan laboratory is gratefully acknowledged.

ABSTRACT

Background: Clinical trial simulations and pharmacometric modeling of biomarker profiles for underrepresented groups are challenging because the underlying studies frequently do not have sufficient participants from these groups.

Objectives: To investigate generative adversarial networks (GANs), an artificial intelligence (AI) technology that enables realistic simulations of complex patterns, for modeling clinical biomarker profiles of underrepresented groups.

Methods: GANs consist of generator and discriminator neural networks that operate in tandem. GAN architectures were developed for modeling univariate and joint distributions of a panel of 16 diabetes-relevant biomarkers from the National Health and Nutrition Examination Survey (NHANES), which contains laboratory and clinical biomarker data from a population-based sample of individuals of all ages, racial groups, and ethnicities. Conditional GANs were used to model biomarker profiles for race/ethnicity categories. GAN performance was assessed by comparing GAN outputs to test data.

Results: The biomarkers exhibited non-normal distributions and varied in their bivariate correlation patterns. Univariate distributions were modeled with generator and discriminator neural networks consisting of two dense layers with rectified linear unit-activation. The distributions of GAN-generated biomarkers were similar to the test data distributions. The joint distributions of the biomarker panel in the GAN-generated data were dispersed and overlapped with the joint distribution of the test data as assessed by

three multi-dimensional projection methods. Conditional GANs satisfactorily modeled the joint distribution of the biomarker panel in the Black, Hispanic, White, and "Other" race/ethnicity categories.

Conclusions: GAN are a promising AI approach for generating virtual patient data with realistic biomarker distributions for under-represented race/ethnicity groups.

INTRODUCTION

There is great interest in harnessing the power and versatility of artificial intelligence (AI) methods in clinical pharmacology and pharmacometrics to facilitate modeling and simulation, accelerate drug development, and improve patient outcomes in drug therapy.

Participants in clinical trials overall are not fully representative of the population of patients [1-5]. In 2011, African Americans and Hispanics comprised 12% and 16% of the United States population, respectively, but only 5% and 1% of clinical trial participants [2]. The Food and Drug Administration (FDA) recently issued a guidance to address representation in clinical trials[5] that recommended broadening study eligibility criteria and addressing other study design and recruitment logistic factors to improve the participant pools. There is clearly an unmet need and knowledge gap to enable modeling of treatment effects and safety in diverse populations and in under-represented groups.

We posit that when successful, the use of AI technologies in the context of already available "big data" could be a transformative computational strategy to mitigate the impact of under-representation of race/ethnicity groups in pharmacometrics and drug development. The clinical pharmacology and pharmacometrics research community has not meaningfully leveraged potentially promising AI-enhanced approaches to address treatment variability, group-level outcome disparities, and real world clinical applications [6]. The increasing availability of public health databases, de-identified health records and pharmacogenomics data provides new opportunities for forecasting of biomarker profiles and drug outcomes in diverse and in under-represented populations [7].

Generative adversarial networks (GANs) are a powerful, deep learning (DL)-based, AI technology that enables realistic simulations of complex patterns [8]. GANs utilize two neural networks called the generator and discriminator that learn from data to generate complex high-dimensional pattern distributions. The generator neural network synthesizes instances of plausible samples that conform to the population distribution of interest. The discriminator network is a classifier that attempts to categorize whether each instance presented as input belongs to the training data or has been synthesized by the generator. The adversarial training process results in a generator that produces samples that mimic the training data.

We hypothesized that GANs could be an effective approach for generating realistic biomarker profiles for clinical trial and pharmacometrics simulations. The objective was to design and assess GANs capable of generating disease-relevant biomarker profiles by learning from observations in real-world, population-based studies. An additional goal was to extend the GAN approach for generating biomarker joint distributions for under-represented race/ethnicity groups.

METHODS

Overview of Studies

Three studies were designed to challenge the utility and assess the performance of GANs for one-dimensional, higher-dimensional, and conditional higher dimensional biomarker distributions.

Univariate Biomarker Distribution Simulations

Dataset: The National Center for Health Statistics (NCHS) conducts an annual survey that assesses the health and nutritional status of adults and children in the United States by means of laboratory measurements, physical screening, and surveys, which are released to the public in biannual cycles [9]. Here, we obtained and pooled the NHANES data from 2009-2010, 2011-2012, 2013-2014, 2015-2016 and 2017-2018 cycles.

We identified a set of 16 diverse diabetes-relevant biomarkers for investigating the utility of GANs for modeling high-dimensional biomarker joint distributions. The following biomarkers were selected: urine creatinine, fasting glucose, insulin, body mass index, glycohemoglobin, triglyceride, total cholesterol, alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma glutamyl transferase (GGT), uric acid, high sensitivity C-reactive protein, direct HDL-cholesterol, average systolic blood pressure, and ferritin. Age, sex, race/ethnicity were obtained as demographic descriptors.

Data Pre-processing: Average systolic blood pressure was a derived variable calculated as the average of 3 systolic blood pressure readings for those with [?] 3 readings and on 2 readings for those with only 2 readings.

The biomarker data were log-transformed, and min-max scaled to the range$[-1, 1]$. The pooled data were randomly split into training (80%) and test (20%) data sets. Listwise exclusion was employed.

GAN Architecture: A common generator neural network architecture was used for modeling all the 16 univariate biomarker distributions, i.e., for the 1-dimensional case.

The generator takes input data from a 10-dimensional latent space that was trained to create output data resembling the training data distribution. The generator neural network model consisted of two dense layers. The hidden layers were comprised of a rectified linear unit (ReLU) function [10] and a batch normalization layer[11]. The batch normalization layers standardize the inputs to the dense layer for each batch and stabilize the learning process [11]. The final layer of the generator was tanh activated.

The discriminator takes input from the generator and predicts whether it belongs to the training distribution. The discriminator model contained two dense layers with ReLu activation. The output was passed to a sigmoid activation function to obtain a classification score. The discriminator network is trained using a binary cross-entropy loss.

Our GANs were prototyped using the AI software tools Keras/TensorFlow. Keras is a neural network library that is integrated with TensorFlow, the open-source library for AI and machine learning.

Training was conducted for 5000 epochs or until the Kolmogorov-Smirnov test p -value for the training vs. generated sample distributions was > 0.05.

Data Analysis: GAN performance was assessed by comparing the GAN-generated biomarker distributions to the test data. For visualization of each biomarker distribution, density histograms containing a sample of 1000 generated and test data samples were used. Quantile-quantile plots of test data vs. GAN-generated data were also assessed.

High-Dimensional Biomarker Panel Joint Distribution Simulations

We developed and evaluated GAN for higher dimensional distributions.

Dataset and Data Pre-Processing: For these experiments, the joint distribution of 14 of the 16 diabetes-relevant biomarkers from the univariate setting was assessed.

High-sensitivity C-reactive protein (hs-CRP) and ferritin were excluded from the list of biomarkers; ferritin was excluded because of sample size and hs-CRP was excluded because assay methodologies changed across the NHANES data sets.

GAN Architecture: The architecture of the conditional GAN was based on Xu *et al* . [12] for tabular data.

Two fully connected hidden layers of size 256 were used in both generator and discriminator. In the generator, batch-normalization and ReLU activation functions were used after each fully connected layer. A variational Gaussian mixture model was used to identify the modality of the data and apply normalization specific to the mode. After two hidden layers, the synthetic row representation is generated. The scalar values of this representation are generated using tanh activation, while the mode indicator and discrete values are generated by Gumbel softmax.

In the discriminator, we used leaky ReLU function and dropout on each hidden layer. The PacGAN framework with 10 samples in each pack was used to reduce mode collapse [13].

The model was trained for 1000 epochs with batch size of 300 and five discriminator steps.

Data Analysis: For visualization, the t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP) and principal components analysis (PCA) were used to obtain the two-dimensional projections of the 14-dimensional data. The*Rtsne* , *umap* packages and *prcomp* function in R were used. The perplexity and theta hyperparameters were set to 50 and 0.5, respectively, for t-SNE. The *ggpairs* package was used to generate pairs panel plots containing univariate densities, bivariate scatter plots and Spearman rank correlation of the test data and GAN-generated distributions. Seven of the 14 biomarkers were assessed in pairs panel plots to keep the number and size of the bivariate plots amenable for visual interpretation.

Conditional GANs for Biomarker Distribution Simulations for Under-represented Groups

Dataset: Conditional GAN analyses were conducted with the same 14-biomarker diabetes-relevant set and test-training methods of the previous High-Dimensional Biomarker Joint Distribution Simulations section.

Data Pre-processing: The race variable was obtained from the*RIDRETH1* variable in the NHANES datasets. The Non-Hispanic Black group was categorized as Black, the Mexican American and Other Hispanic groups were categorized as Hispanic, the Non-Hispanic White group was categorized as White, and the Other Race-Including Multi-Racial was categorized as Other.

GAN Architecture: The generator and discriminator architectures were identical to that used for High-Dimensional Biomarker Panel Joint Distribution Simulations. However, the derived race/ethnicity categories were encoded as one-hot encoded vectors and appended with the biomarker input.

The model was trained for 1000 epochs with batch size of 300 and five discriminator steps.

Data Analysis: The high dimensional distributions were visualized using t-SNE and UMAP methods and assessments of the univariate distribution of the GAN-generated distribution vs. test data distribution were conducted with box plots.

RESULTS

Overview of Generative Adversarial Networks

Generative Adversarial Networks (GANs) are an artificial intelligence (AI) method for simulating samples from complex data distributions. GANs cleverly harness neural networks, which are powerful and versatile at learning approximations for arbitrary high dimensional functions.

A GAN is a system with two neural networks, the generator and discriminator (Figure 1), which interact in a mutually competitive (adversarial) supervised learning strategy. Samples (training data) from the target data distribution serve as input for training the GAN. The output is a learned model representing the underlying data distribution.

The input to the generator is a random variable vector drawn from a latent space that provides a supply of random variables of suitable dimensionality. The generator neural network transforms the input from the latent space to synthesize generated data as output.

The inputs to the discriminator are instances of generated data and training data. The discriminator neural network is a binary classifier that is designed to determine whether a given input was drawn from the training or was generated data. The classification errors are used to compute the generator and discriminator loss functions.

Backpropagation of the loss functions is used to update the parameters of the generator and discriminator neural networks via gradient descent to enable supervised learning. The training process in a GAN is adversarial because the generator is trained to maximize the classification error, while the discriminator is

trained to minimize the classification error. This adversarial strategy guides the generator neural network to become increasingly proficient at synthesizing data that approximates the training data distribution. Ideally after training is complete, the generated output is a random variable indistinguishable from the target data distribution. The performance of the GAN is usually assessed on independent test data.

The GAN approach was used to simulate high-dimensional joint distributions of disease-relevant biomarkers of virtual patient populations for pharmacometrics applications. Large heterogeneous public domain datasets with biomedical information on diverse populations were utilized for GAN training and performance evaluation.

Univariate Biomarker Distribution Simulations

As a first step, we sought proof-of-concept evidence to motivate the use of GAN for pharmacometrics. We focused on modeling the univariate distributions of 16 biomarkers that are clinically relevant for diabetes.

Table 1 summarizes demographic characteristics and the biomarker levels in the data set.

Figure 2 compares the distribution of the biomarkers in the training set to the generated distribution from the GANs for eight biomarkers; the remaining eight biomarkers are summarized in Supplementary Figure 1. Despite the log transformation, the set of scaled distributions for the biomarkers had diversity of patterns and evidence for non-normality: e.g., some of the biomarkers were left skewed (e.g., urine creatinine, Figure 2B), some were right skewed (e.g., fasting glucose, Figure 2C) and some had broad distributions (e.g., body mass index, Figure 2E and high sensitivity C-reactive protein, Supplementary Figure 1M).

The dark gray regions of the histograms show the overlap of the generated density histograms (salmon) and the test data density histograms (teal). The extensive regions of overlap in Figure 2 and Supplementary Figure 1 indicate the satisfactory concordance of GAN-generated distributions to the test data distribution for the 16 biomarkers.

The concordance was further assessed using quantile-quantile plots and Kolmogorov-Smirnov tests (Supplementary Figure 2). The quantile-quantile plots showed extensive clustering around the line of identity. The p -values from the Kolmogorov-Smirnov test were not significant ($p > 0.05$) for the majority of GAN-generated biomarkers distributions. However, GAN-generated distributions for glucose, aspartate aminotransferase, gamma glutamyl transferase and high sensitivity C-reactive protein had p [?] 0.05 despite the overall visual similarity with the test histogram probability distribution function.

These promising proof-of-concept results motivated further, more rigorous investigation of GAN applications for scenarios relevant to drug development and pharmacometrics.

High-Dimensional Biomarker Panel Joint Distribution Simulations

We evaluated whether GANs could be used to generate the joint distribution of multiple biomarkers by using the 14 diabetes-relevant biomarkers.

Because the 14-dimensional joint distribution is not amenable to visualization, we used three different multi-dimensional visualization approaches, t-SNE (Figure 3A), UMAP (Figure 3B), and PCA (Figure 3C) to generate 2-dimensional projections of the test and GAN-generated distributions. The projected data for the GAN-generated distribution (teal circles) was well dispersed in the test data distribution (salmon circles) for all three approaches. This indicates that GANs are a promising approach for generating high dimensional biomarker distributions.

To further assess the performance of GANs, we visualized the univariate and bivariate marginal distributions from the high dimensional joint distribution (Figure 3D) using pairs panel plots, which summarize the univariate density along the diagonal, the bivariate scatter plots in the lower triangular region and the Spearman correlation coefficients in the upper triangular region. The pairs panel plots for scaled log-transformed levels of seven biomarkers: urine albumin, urine creatinine, fasting glucose, insulin, body mass index, glycohemoglobin and triglyceride are shown in Figure 3D. The univariate densities (see diagonal in Figure 3D) for

the GAN-generated data for all seven biomarkers overlapped extensively with the test data density and the individual density curves were difficult to distinguish. The bivariate scatter plots also overlapped extensively, and the GAN-generated data points were evenly dispersed among the test data points for all 21 bivariate plots in Figure 3D.

These results show that GAN-generated distributions can be useful for modeling systems of clinical biomarkers.

Conditional Biomarker Distribution Simulations for Under-represented Groups

A conditional GAN was used to evaluate whether the GAN method could be used to generate biomarker distributions in Black, Hispanic, Other and White under-represented minority groups.

The number (%) of Black, Hispanic, Other and White subjects in the test data set were 1730 (20.8%), 2228 (26.8%), 1137 (13.7%) and 3230 (38.8%); the total number of subjects was 8325.

The t-SNE projections of the GAN-generated data and the test data distributions for the four race categories are compared in Figure 4. The corresponding UMAP projections are summarized in Supplementary Figure 3. The t-SNE and UMAP projections for the GAN-generated distributions were qualitatively well-dispersed across the test data for the four race/ethinicity groups. The box plots in Figure 5 and Supplementary Figure 4 compare the univariate distributions of the 14 biomarkers and demonstrate the concordance of the GAN-generated data with the test data for each race.

Together, these results demonstrate that the GAN strategy can generate satisfactory approximations for high dimensional biomarker joint distributions in under-represented groups.

DISCUSSION

The GAN approach investigated integrates AI techniques with the large public domain NHANES database containing biomedical information on diverse populations that could prove valuable in pharmacometrics applications. Proof-of-concept computational experiments were conducted to evaluate the capabilities of GANs to simulate univariate distributions of a test bed of 16 diabetes-relevant biomarkers. In the next step, the GAN strategy was extended to complex joint distributions of multiple biomarkers and finally, a conditional GAN was used for modeling of Black, Hispanic, Other and White race/ethnicity categories. The training-test strategy was used for GAN performance evaluation.

The GAN strategy enables robust learning and can be considered "non-parametric" because it does not need prior distributions, which are required for Bayesian approaches. While the latent space for a GAN generator is sampled from a multivariate Gaussian distribution, it serves only as a source of random noise for the generator neural network to transform. Notably, the GAN architecture is indirect because it does not conduct head-to-head comparison of the generated data distribution vs. training data distribution. GANs avert direct comparison by intercalating a binary classifier and judicious use of the adversarial loss functions. The literature on GANs in pharmacometrics is sparse. Parikh *et al* . [14] have used GANs to generate instances of models for cardiac mechanics in control myocytes and myocytes treated with omecamtiv mecarbil, a new drug for treating heart failure. The GANs were used to find model parameters for fitting the data for both groups. This application of GANs to *in vitro* data differs qualitatively from the patient-centric problem in our research.

Conditional GANs are an extension of GANs wherein the generator and discriminator networks are conditioned with additional input. Conditional GANs are particularly useful for modeling multimodal data and have been used elsewhere for tagging and annotating images[15]. We found that biomarker profile joint distribution could be modeled using GAN architectures effective for tabular data, which can consist of multiple data types, e.g., continuous variables, ordinal, and categorical. Tabular data generation presents some unique challenges as compared to GAN modeling of images because: i) columns in a row do not have local structure and, ii) conditioned variable-dependent continuous variables are generally multimodal (i.e., the density function has several peaks). The typical GAN architectures designed for images are not particu-

larly good at generating multimodal data because of a phenomenon termed "mode collapse". Mode collapse reduces the diversity of output samples and occurs when the generator can only produce a single type of output or a small set of outputs that fool the discriminator [13]. To simultaneously generate a mix of discrete and continuous columns, the Xu *et al* .[12] GAN approach applies both softmax and tanh on the output. We used the PacGAN method, wherein the discriminator decision-making is guided by multiple or "packed" samples from each class [13]. In PacGAN, the discriminator does not classify each generated sample but instead, examines a "pack" of samples for a class. Thus, diversity of the generated samples becomes a criterion for the discriminator in the classification process and helps avoid mode collapse. By implementing these enhancements[12,13], we found that a conditional GAN yielded effective results for modeling race/ethnicity. The approach addresses the frequency differences between the various under-represented groups, and the multimodality resulting from between-group differences in biomarker expression.

We selected 16 diverse diabetes-relevant physiological biomarkers that reflected different organ systems and become clinically salient at different stages of diabetes progression. Alterations to plasma glucose and insulin profiles are direct consequences of diabetes and can be dysregulated early in diabetes because of decreased pancreatic β-cell function or increased insulin resistance in hepatic and peripheral tissues. Glycohemoglobin is related to the average glucose exposure over 2-3 months. In contrast, increased urinary creatinine and albumin are the result of compromised renal function during diabetes disease progression. We also included integrative biomarkers, e.g., body mass index and systolic blood pressure, metabolic biomarkers, e.g., triglycerides and cholesterol, inflammatory biomarkers (C-reactive protein and ferritin) and hepatic biomarkers (e.g., alanine aminotransferase, aspartate aminotransferase and gamma glutamyltransferase) that are dysregulated in diabetes.

One of the strengths of the NHANES as a source of "big data" for modeling under-represented groups is that while the total sample size in a given cycle is fixed, the survey adapts its population-based sampling strategy to include adequate numbers of individuals from under-represented groups, e.g., there is ongoing oversampling of Hispanics, non-Hispanic Blacks, older adults, and low income whites/others groups and beginning in 2011, non-Hispanic Asians were oversampled [16]. We used the *RIDRETH1* variable from NHANES to derive our under-represented groups; additional race-ethnicity variables have been added to NHANES, but these variables were not available across all the datasets we used. A weakness is that the NHANES sample is limited to the non-institutionalized civilian resident population: it does not contain groups such as prisoners, military personnel, individuals in psychiatric institutions, and drug rehabilitation facilities. Interestingly, Allen *et al* . and Rieger*et al* . also leveraged NHANES data in their work on virtual patients [17,18]. We have previously used NHANES as the data source in the generalized pharmacometrics modeling (GPM) approach, which integrates population models with AI techniques. GPM simulates pharmacokinetic (PK) parameters from population PK covariate models using Bayesian networks that include demographic and biomarker features identified from NHANES. The integration of external data enables GPM to facilitate modeling and simulation of drug disposition and effects for populations different from those in the underlying PK study[7].

Creating virtual populations requires modeling or otherwise sampling the joint distribution of biomarkers of interest. If the biomarkers are not normally distributed or if there are multiple biomarkers of interest, covariance matrices are generally inadequate for characterizing higher-order inter-dependencies. General empirically-motivated methods for producing virtual patient populations include patient selection using inclusion and exclusion criteria [19], bootstrapping similar clinical trials or patient databases[20] and simulating from fitted distributions[21]. Simulated annealing and nested simulated annealing-based methods have been proposed for generating "plausible" populations in the context of quantitative systems pharmacology models[17,18]. Our GAN approach relies on neural network-based learning and is generative, i.e., it creates new sample sets: it differs substantially from the non-parametric re-sampling and parametric Bayesian approaches that have been used in pharmacometrics for approximating data distributions.

GANs are considered a deep learning (DL) method as many GANs require deep neural networks (DNN; "deep" refers to the number of network layers) for the generator and discriminator architectures. Although there is increasing interest in leveraging AI approaches including DL in drug discovery and development, the

assessments of DL and GANs in pharmacometrics have been preliminary [22,23]. Liu*et al* . [23] used long short-term memory (LSTM, a common neural network architecture that is effective for time series) DNN to model simulated PK/PD data of a hypothetical drug. The plasma concentration and effect level under one dosing regimen was used to train the model and the model was used to predict the individual PK/PD for other dosing regimens. Lu [22] included neural ordinary differential equations for forecasting PK/PD of platelet responses in a clinical dataset of 800 patients. It should be noted that like many AI and DL methods, GAN methods can be computationally intensive; however, graphic processing units (GPU) and high-performance computing (HPC) architectures can improve the performance of AI algorithms substantially [24,25].

Our results demonstrate the potential of the GAN approach for modeling the joint distribution of complex systems of disease-relevant biomarkers in under-represented groups. The approach may find utility for generating virtual patient populations for clinical trial simulations and pharmacometrics.

CONFLICT OF INTEREST DISCLOSURE

Rahul Nair, Sandra Frank, and Deen Dayal Mohan have no conflicts.

Srirangaraj Setlur and Venu Govindaraju received unrelated research funding from the National Science Foundation, United State Postal Service, and the Intelligence Advanced Research Projects Activity agencies. received unrelated research funding from the National Science Foundation, United States Postal Service, and the Intelligence Advanced Research Projects Activity agencies.

Murali Ramanathan received research funding from the National Science Foundation, Otsuka Pharmaceuticals, and the National Institutes of Health.

REFERENCES

1. Loree JM, Anand S, Dasari A, Unger JM, Gothwal A, Ellis LM, Varadhachary G, Kopetz S, Overman MJ, Raghav K. Disparity of Race Reporting and Representation in Clinical Trials Leading to Cancer Drug Approvals From 2008 to 2018. *JAMA Oncol.* Oct 1 2019;5(10):e191870.

2. Clark LT, Watkins L, Pina IL, Elmer M, Akinboboye O, Gorham M, Jamerson B, McCullough C, Pierre C, Polis AB, Puckrein G, Regnante JM. Increasing Diversity in Clinical Trials: Overcoming Critical Barriers. *Curr Probl Cardiol.* May 2019;44(5):148-172.

3. Lee M. We Must Act Now: Building Trust and Increasing Minority Participation in COVID-19 Clinical Trials. *Dela J Public Health.* Nov 2020;6(5):34-35.

4. Webber-Ritchey KJ, Aquino E, Ponder TN, Lattner C, Soco C, Spurlark R, Simonovich SD. Recruitment Strategies to Optimize Participation by Diverse Populations. *Nurs Sci Q.* Jul 2021;34(3):235-243.

5. Center for Biologics Evaluation and Research, Center for Drug Evaluation and Research. Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry. Silver Spring, MD: Food and Drug Administration; 2020.

6. D'Argenio V. The high-throughput analyses era: Are we ready for the data struggle? *high-throughput.* 2018;7(8):1-12.

7. McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. *Br J Clin Pharmacol.* Feb 2022;88(4):1482-1499.

8. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks.*arXiv.* 2014:arXiv:1406.2661 [stat.ML].

9. National Health and Nutrition Examination Survey. About the National Health and Nutrition Examination Survey. Hyattsville, MD: National Center for Health Statistics; 2017.

10. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. Proceedings of the 27th International Conference on Machine Learning; 2010; Haifa, Israel.

11. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of Machine Learning Research (PMLR); 2015/06/01.

12. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019; Vancouver, Canada.

13. Lin Z, Khetan A, Fanti G, Oh S. PacGAN: The power of two samples in generative adversarial networks. *arXiv.*2017:arXiv:1712.04086

14. Parikh J, Rumbell T, Butova X, Myachina T, Acero JC, Khamzin S, Solovyova O, Kozloski J, Khokhlova A, Gurev V. Generative adversarial networks for construction of virtual populations of mechanistic models: simulations to study Omecamtiv Mecarbil action.*J Pharmacokinet Pharmacodyn.* Oct 29 2021.

15. Mirza M, Osindero S. Conditional Generative Adversarial Nets. *arXiv.* 2014:arXiv:1411.1784 [cs.LG].

16. National Health and Nutrition Examination Survey. National Health and Nutrition Examination Survey: NHANES 2015-2016 Overview. In: National Center for Health Statistics, ed: Centers for Disease Control; 2015.

17. Allen RJ, Rieger TR, Musante CJ. Efficient Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models. *CPT Pharmacometrics Syst Pharmacol.* Mar 2016;5(3):140-146.

18. Rieger TR, Allen RJ, Bystricky L, Chen Y, Colopy GW, Cui Y, Gonzalez A, Liu Y, White RD, Everett RA, Banks HT, Musante CJ. Improving the generation and selection of virtual populations in quantitative systems pharmacology models. *Prog Biophys Mol Biol.* Jun 15 2018.

19. Kimko HC, Duffull SB. *Simulation for designing clinical trials: a pharmacokinetic-pharmacodynamic modeling perspective.* New York: Marcel Dekker; 2003.

20. Goldenholz DM, Tharayil J, Moss R, Myers E, Theodore WH. Monte Carlo simulations of randomized clinical trials in epilepsy.*Ann Clin Transl Neurol.* Aug 2017;4(8):544-552.

21. Brainard J, Burmaster DE. Bivariate distributions for height and weight of men and women in the United States. *Risk Anal.* Jun 1992;12(2):267-275.

22. Lu J, Bender B, Jin JY, Guan Y. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature Machine Intelligence.* 2021/08/01 2021;3(8):696-704.

23. Liu X, Liu C, Huang R, Zhu H, Liu Q, Mitra S, Wang Y. Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling. *Int J Clin Pharmacol Ther.* Feb 2021;59(2):138-146.

24. Mittal S, Vaishay S. A survey of techniques for optimizing deep learning on GPUs. *Journal of Systems Architecture.*2019/10/01/ 2019;99:101635.

25. Oh K-S, Jung K. GPU implementation of neural networks.*Pattern Recognition.* 2004/06/01/ 2004;37(6):1311-1314.

TABLES

Table 1. Detailed summary statistics of the demographics and biomarkers from data set combined from NHANES 2009-2010, 2011-2012, 2013-2014, 2015-2016 and 2017-2018. The total number of cases was$n$ = 29,547 cases.

Variable Name	Variable Description	Actual N (% Missing)	Percent
RIAGENDR	Sex Female Male	29547 (0) 14909 (0) 14638 (0)	– 50.5 49.5

Variable Name	Variable Description	Actual N (% Missing)	Percent		
RIDETH1	Race/Ethnicity Non-Hispanic Black Mexican American Other Hispanic Non-Hispanic White Other–Including multi-racial	29547 (0) 6755 5106 3029 10543 4114	– 22.9% 17.3% 10.3% 35.7% 13.9%		
			Mean (SD)	**Median (IQR)**	**Min – Max**
RIDAGEYR	Age, years	29547(0)	32.74(25)	29(10, 54)	0.0 - 80*
URXUMA	Albumin, urine (μg/ml)	23666 (19.9)	41 (268)	8.4 (4.4, 17.4)	0.21 - 14800
URXUCR	Creatinine, urine (mg/dl)	23666 (19.9)	123 (81.6)	109 (62, 167)	3 - 800
LBXGLU	Fasting glucose (mg/dl)	9310 (68.4)	107.2 (32.6)	99 (92, 109)	36 - 451
LBXIN	Insulin (μU/ml)	9043 (69.3)	14.7 (17.4)	10.7 (6.8, 17.2)	0.14 - 647
BMXBMI	Body mass index (kg/m^2)	26019 (11.9)	25.9 (7.91)	25.2 (19.9, 30.4)	12.3 - 86.19
LBXGH	Glycohemoglobin (%)	19120 (35.3)	5.7 (1.02)	5.5 (5.2, 5.8)	3.6 - 17.7
LBXTR	Triglyceride (mg/dl)	9181 (68.9)	117.1 (99.04)	95 (65, 140)	10 - 2742
LBXTC	Total cholesterol (mg/dl)	21572 (26.9)	183.05 (41.2)	179 (59, 208)	59 - 528
LBXSATSI	Alanine aminotransferase (ALT) (IU/L)	18733 (36.5)	23.4 (20.06)	19 (14, 26)	2 - 1363
LBXSASSI	Aspartate aminotransferase (AST) (IU/L)	18709 (36.6)	24.4 (15.1)	22 (18, 27)	6 - 733
LBXSGTSI	Gamma glutamyl transferase (GGT) (IU/L)	18735 (36.5)	27.1 (37.6)	18 (13, 28)	2 - 1192
LBXSUA	Uric acid (mg/dl)	18734 (36.5)	5.3 (1.4)	5.3 (4.3, 6.3)	0.4 - 15.1
LBXHSCRP	HS C-Reactive Protein (mg/L)	15549 (47.3)	1.79 (5.3)	0.42 (0.11, 1.42)	0.01 - 182.8

Variable Name	Variable Description	Actual N (% Missing)	Percent		
LBDHDD	Direct HDL-cholesterol (mg/dl)	21573 (26.9)	52.8 (14.7)	51 (42, 61)	10 – 189
AVGSBP#	Average systolic blood pressure, mmHg	21245 (28.1)	119.5 (19.02)	116 (106, 129)	72.6 - 234.6
LBXFER	Ferritin (ng/ml)	2909 (90.1)	47 (54)	32 (19, 57)	2.0 - 1090

* RIDAGEYR is age in years and subjects 80 years-old and over are coded as 80 years.

#AVGSBP is a derived variable calculated as average of 3 systolic blood pressure readings for those with [?] 3 readings and on 2 readings for those with only 2 readings.

FIGURE LEGENDS

Figure 1. Schematic of the generative adversarial network (GAN) method. A GAN consists of two neural networks: the generator and the discriminator. The generator takes random variables from a latent space as input and computes generated data via its neural network. The discriminator takes the training data containing biomarkers and the generated data from the generator as inputs. The neural network in the discriminator is a binary classifier that computes the generator and discriminator loss functions that are used to update the generator and the discriminator neural networks via back propagation.

Figure 2. Figure 2 compares the probability density histogram of a representative generated data set from the generative adversarial network (teal bars) to the probability density histogram of the test data (salmon bars) from the univariate analyses of 8 diabetes-associated biomarkers. The dark gray bars correspond to the regions of overlap between the two probability density histograms. Eight biomarkers are shown: urine albumin (Figure 2A), urine creatinine (Figure 2B), fasting glucose (Figure 2C), insulin (Figure 2D), body mass index (Figure 2E), glycohemoglobin (Figure 2F), triglyceride (Figure 2G), and total cholesterol (Figure 2H). The x -axes on all graphs are biomarker levels that are log-transformed and scaled to lie between -1 and 1. The$p$ -values from the Kolmogorov-Smirnov test are shown on the top left.

Figure 3. The t -stochastic neighbor embedding (t-SNE, Figure 3A), uniform manifold approximation and projection (UMAP, Figure 3B) and principal component analysis two-dimensional projections of the 14-dimensional, diabetes-associated biomarkers data. The test data results are shown in salmon circles and the GAN-generated results are in teal circles. The x -axis (t-SNE X and UMAP X) and y -axis (t-SNE Y and UMAP) correspond to the t-SNE and UMAP projections into two dimensions of the input of 14-dimensional biomarker levels that are log-transformed and scaled to lie between -1 and 1. The PC 1 and PC 2 on the x -axis and y -axis of Figure 3C correspond to the first and second principal components, respectively. Figure 3D is a pairs panel that compares the univariate and bivariate GAN-generated distributions (teal circles) to the test data (salmon circles). The diagonal contains the univariate density for the GAN-generated and test data distributions. The area of overlap is shaded dark gray. The upper triangular region contains the Spearman bivariate correlation coefficients for the test (salmon font) and GAN-generated distributions (teal font). Only 7 of the 14 variables are shown. All variables were log-transformed and scaled to lie in the range [-1, 1]: ALB: Albumin, urine; CRE: Creatinine, urine; GLU: Fasting glucose; INS: Insulin; BMI: Body mass index; GLHB: Glycohemoglobin; TG: Triglyceride.

Figure 4. The t -stochastic neighbor embedding (t-SNE) two-dimensional projections of the 14-dimensional, diabetes-associated biomarkers data for the Black, Hispanic, Other and White race categories. The test data results are shown in salmon circles and the GAN-generated results are in teal circles. The x -axis (t-SNE X)

and y -axis (t-SNE Y) correspond to the t-SNE projections into two dimensions of the input of 14-dimensional biomarker levels that are log-transformed and scaled to lie between -1 and 1.

Figure 5. Box plots of the univariate results from 14-dimensional, diabetes-associated biomarkers data for the Black, Hispanic, Other and White race categories. The test data are shown in salmon, and the GAN-generated results are in teal. The univariate results for eight of 14 diabetes-associated biomarkers are shown: urine albumin (Figure 5A), urine creatinine (Figure 5B), fasting glucose (Figure 5C), insulin (Figure 5D), body mass index (Figure 5E), glycohemoglobin (Figure 5F), triglyceride (Figure 5G), and total cholesterol (Figure 5H). The y -axes on all graphs are biomarker levels that are log-transformed and scaled to lie between -1 and 1. The lines on the box correspond to the 25^{th} quantile, median and 75^{th} quantile, the error bars correspond to the median $\pm$ 1.5 inter-quartile range and the outliers are in black circles.

FIGURE 1

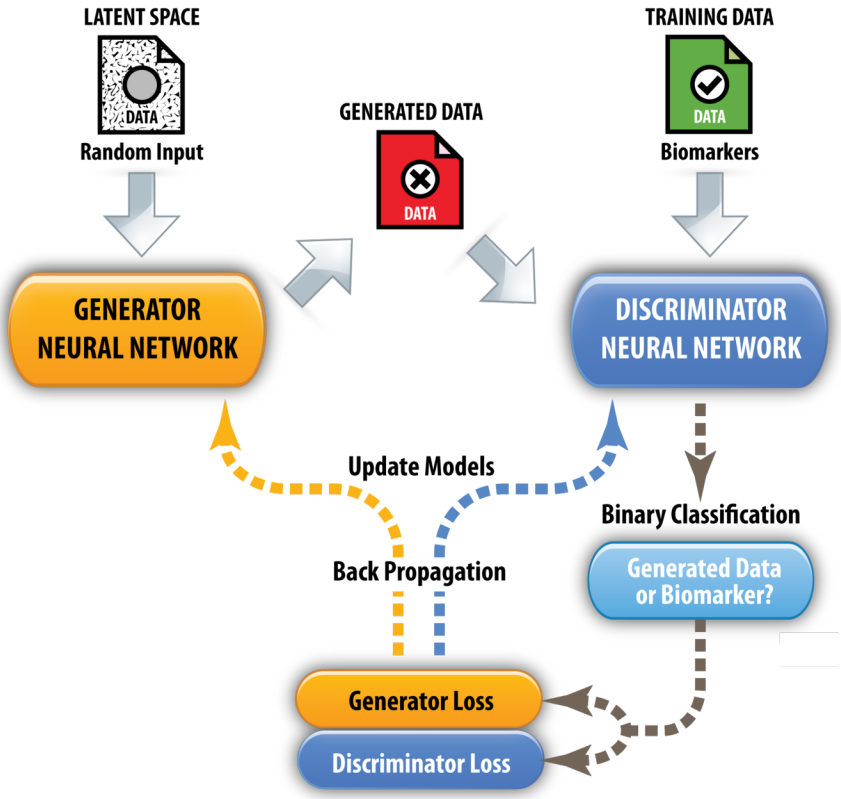


FIGURE 2

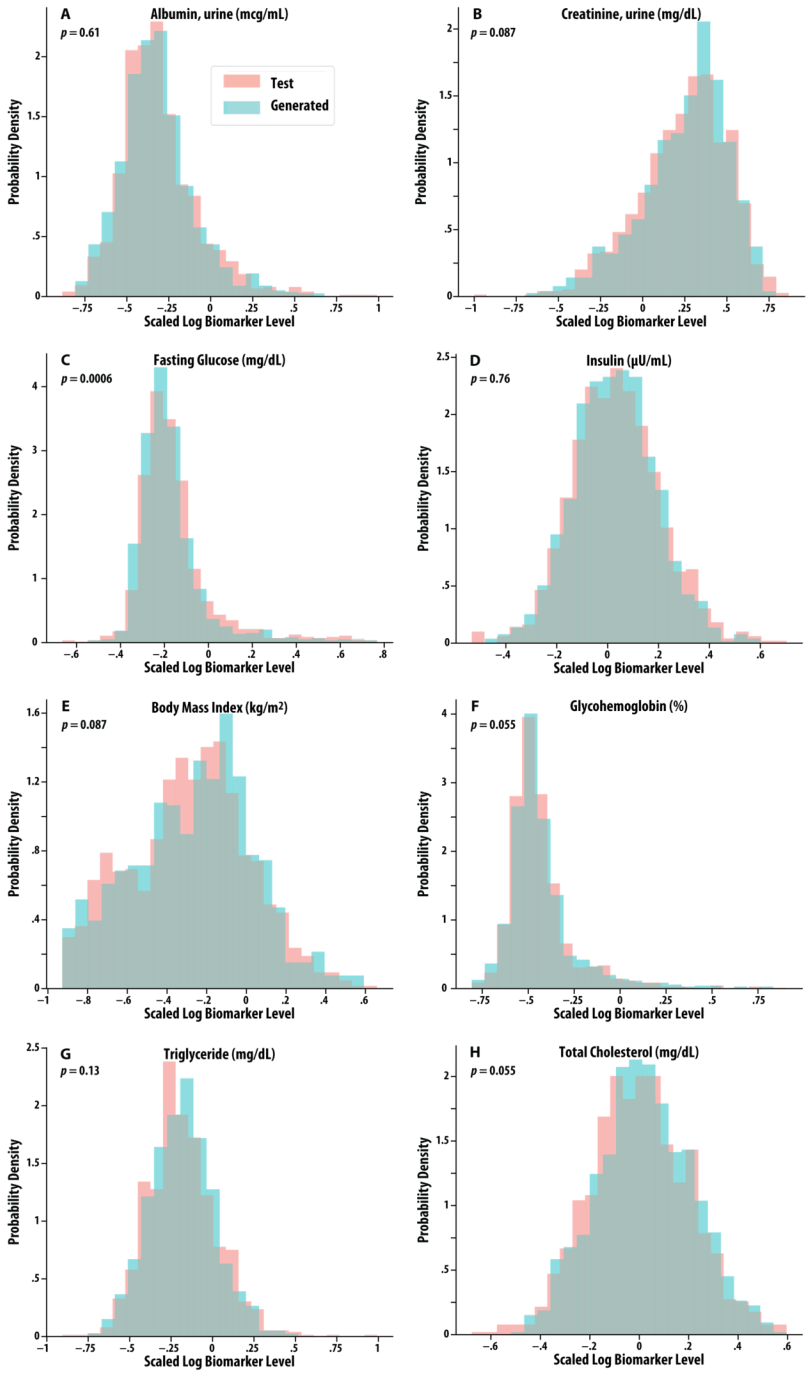


FIGURE 3

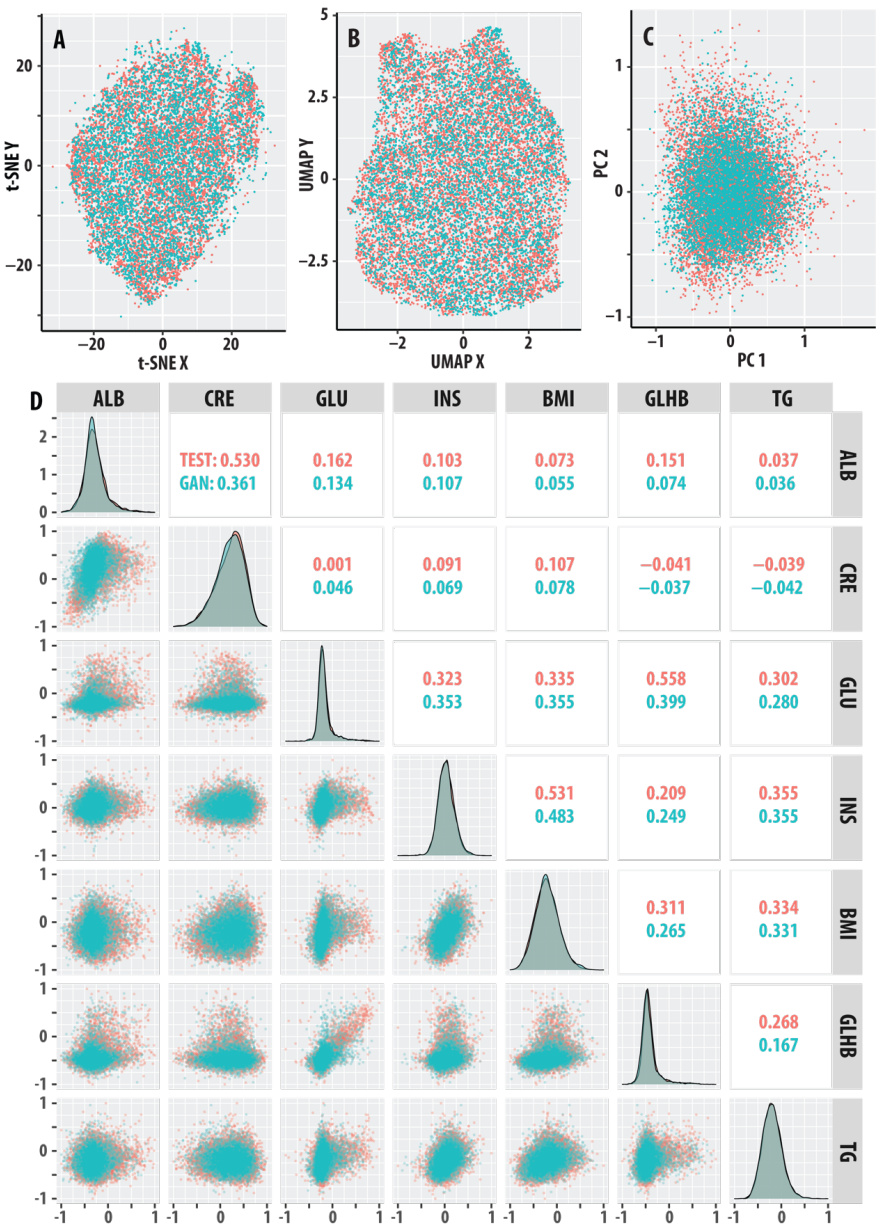


FIGURE 4

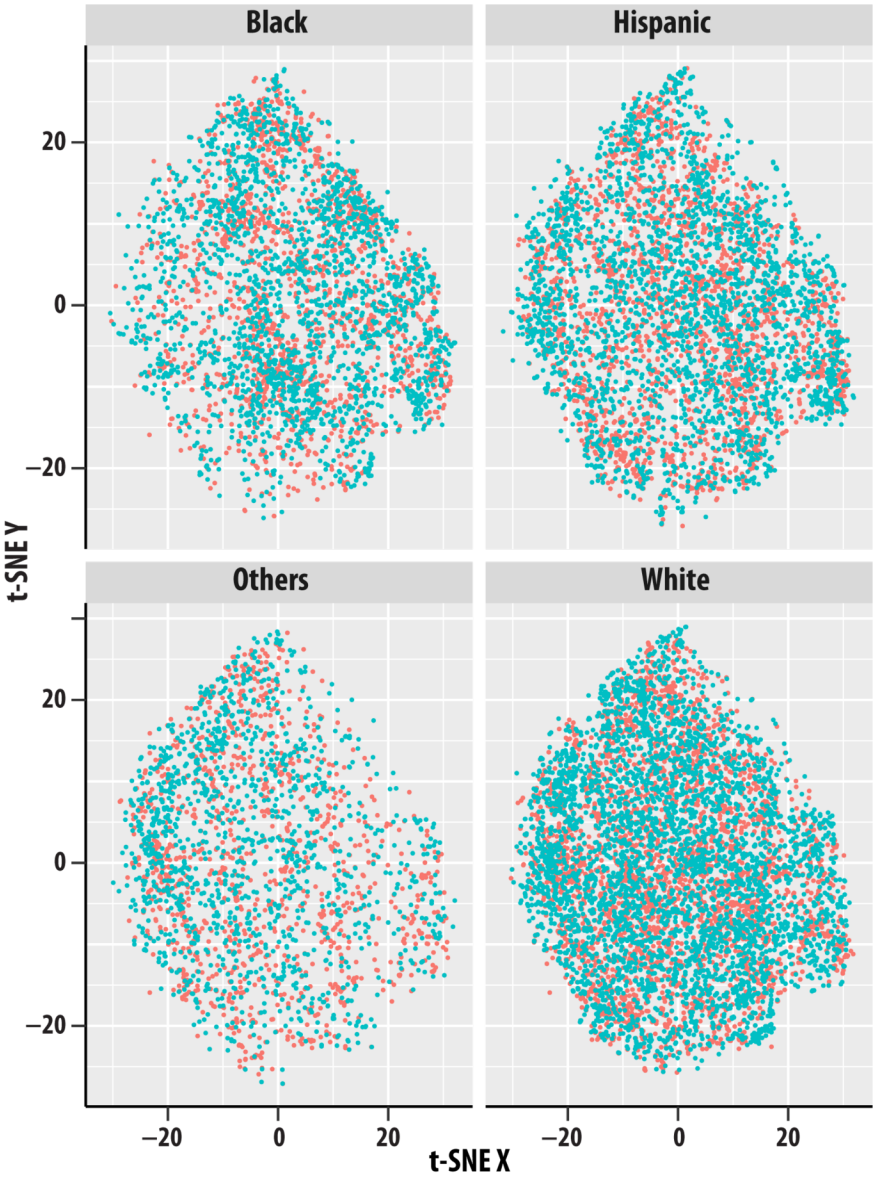
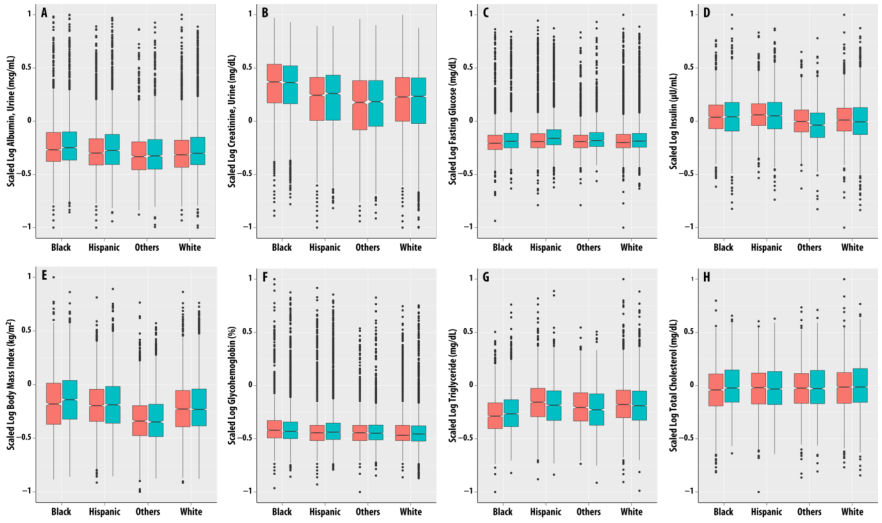


FIGURE 5



SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Supplementary Figure 1 compares the probability density histogram of a representative generated data set from the generative adversarial network (teal bars) to the probability density histogram of the test data (salmon bars) from the univariate analyses of 8 diabetes-associated biomarkers. The dark gray bars correspond to the regions of overlap between the two probability density histograms. The eight biomarkers not shown in Figures 2A-H are shown here: alanine aminotransferase (ALT) (Supplementary Figure 1I), aspartate aminotransferase (AST) (Supplementary Figure 1J), gamma glutamyl transferase (GGT) (Supplementary Figure 1K), uric acid (Supplementary Figure 1L), high sensitivity C-reactive protein (Supplementary Figure 1M), direct HDL-cholesterol (Supplementary Figure 1N), average systolic blood pressure (Supplementary Figure 1O), and ferritin (Supplementary Figure 1P). The x -axes on all graphs are biomarker levels that are log-transformed and scaled to lie between -1 and 1. The p -values from the Kolmogorov-Smirnov test are shown on the top left.

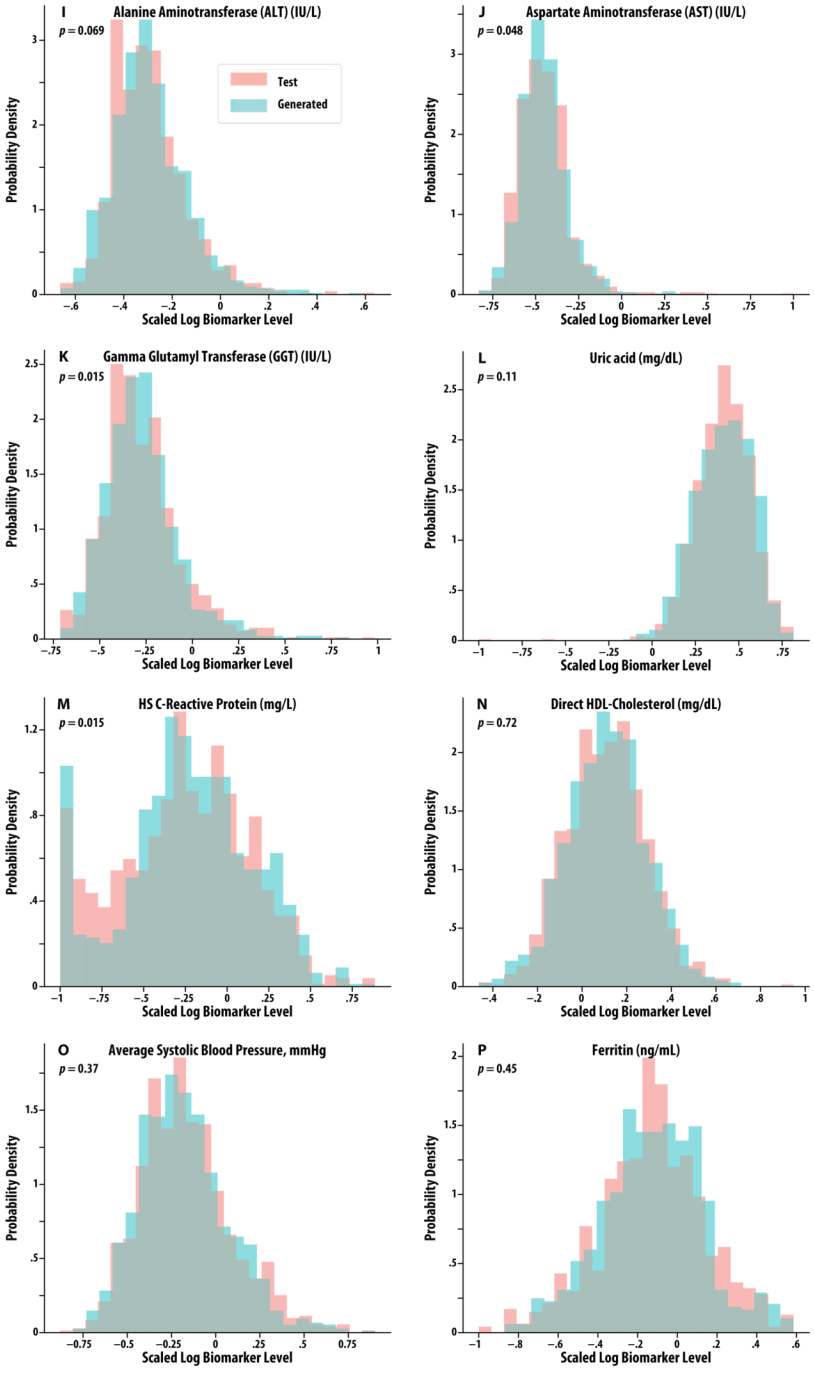
Supplementary Figure 2. Supplementary Figure 2 shows quantile-quantile plots that compare the distribution of a representative generated data set from the generative adversarial network to the distribution of the test data from the univariate analyses of the 16 diabetes-associated biomarkers. The circles correspond to the data and the salmon line is the line of identity. The biomarkers shown are: urine albumin (Supplementary Figure 2A), urine creatinine (Supplementary Figure 2B), fasting glucose (Supplementary Figure 2C), insulin (Supplementary Figure 2D), body mass index (Supplementary Figure 2E), glycohemoglobin (Figure 2F), triglyceride (Figure 2G), total cholesterol (Figure 2H), alanine aminotransferase (ALT) (Supplementary Figure 2I), aspartate aminotransferase (AST) (Supplementary Figure 2J), gamma glutamyl transferase (GGT) (Figure 2K), uric acid (Supplementary Figure 2L), high sensitivity C-reactive protein (Supplementary Figure 2M), direct HDL-cholesterol (Supplementary Figure 2N), average systolic blood pressure (Supplementary Figure 2O), and ferritin (Supplementary Figure 2P).

Supplementary Figure 3. The uniform manifold approximation and projection (UMAP) two-dimensional projections of the 14-dimensional, diabetes-associated biomarkers data for the Black, Hispanic, Other and White race categories. The test data results are shown in salmon, and the GAN-generated results are in teal. The x -axis (UMAP X) and$y$ -axis (UMAP Y) correspond to the UMAP projections into two dimensions of the input of 14-dimensional biomarker levels that are log-transformed and scaled to lie between -1 and 1.

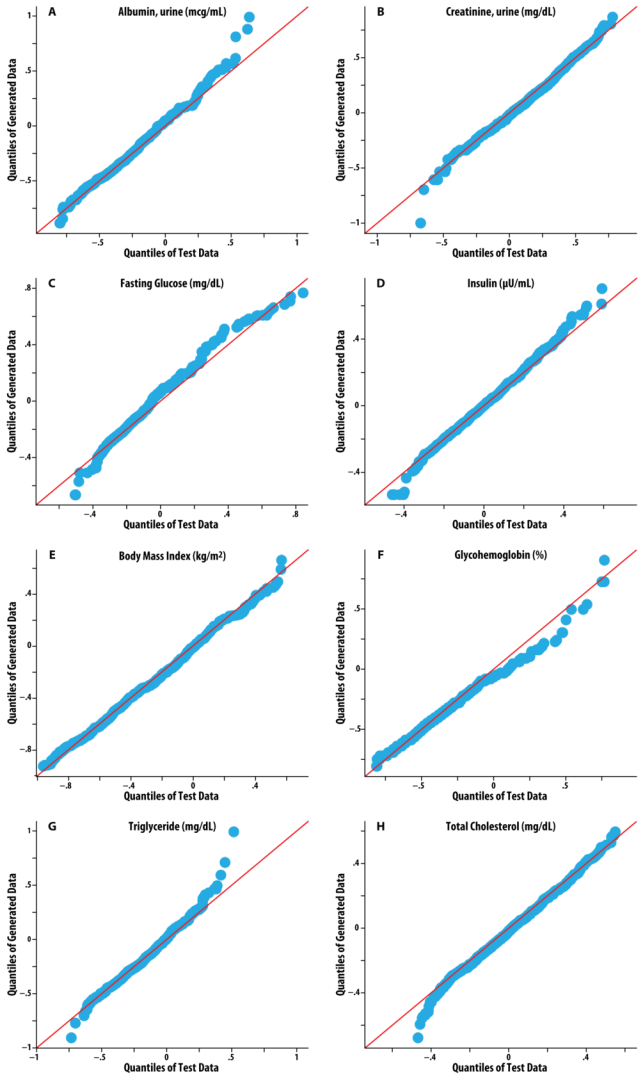
Supplementary Figure 4. Box plots of the univariate results from 14-dimensional, diabetes-associated biomarkers data for the Black, Hispanic, Other and White race categories. The test data are shown in salmon, and the GAN-generated results are in teal. The six biomarkers not shown in Figures 5A-H are shown here: alanine aminotransferase (ALT) (Supplementary Figure 4I), aspartate aminotransferase (AST) (Supplemen-

tary Figure 4J), gamma glutamyl transferase (GGT) (Supplementary Figure 4K), uric acid (Supplementary Figure 4L), direct HDL-cholesterol (Supplementary Figure 4M), and average systolic blood pressure (Supplementary Figure 1N). The x -axes on all graphs are biomarker levels that are log-transformed and scaled to lie between -1 and 1. The lines on the box correspond to the 25^{th}quantile, median and 75^{th} quantile, the error bars correspond to the median $\pm$ 1.5 inter-quartile range and the outliers are in black circles.

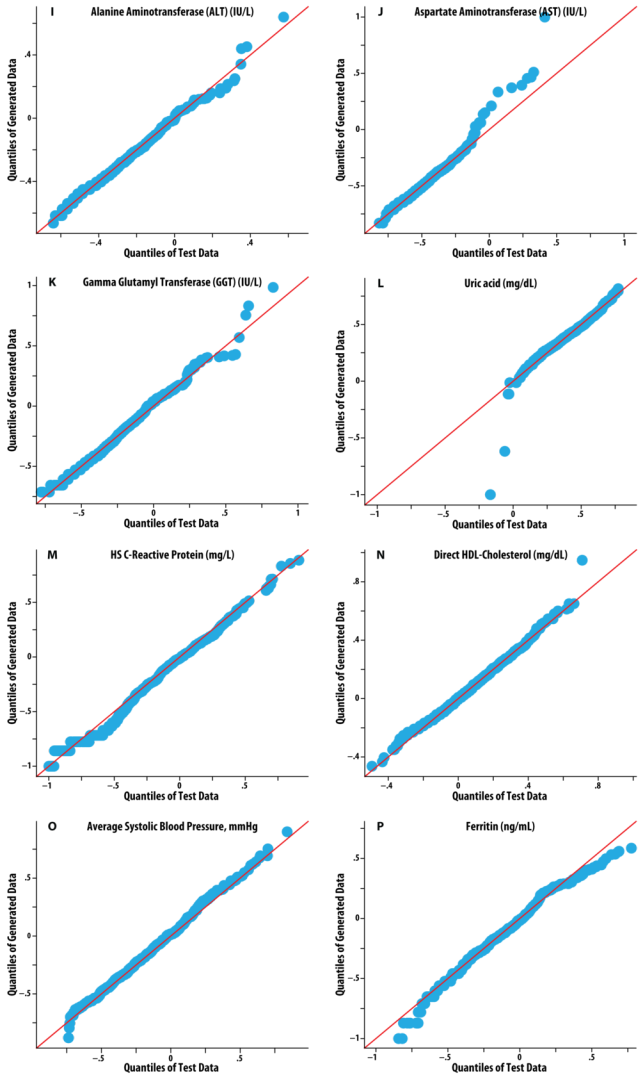
SUPPLEMENTARY FIGURE 1



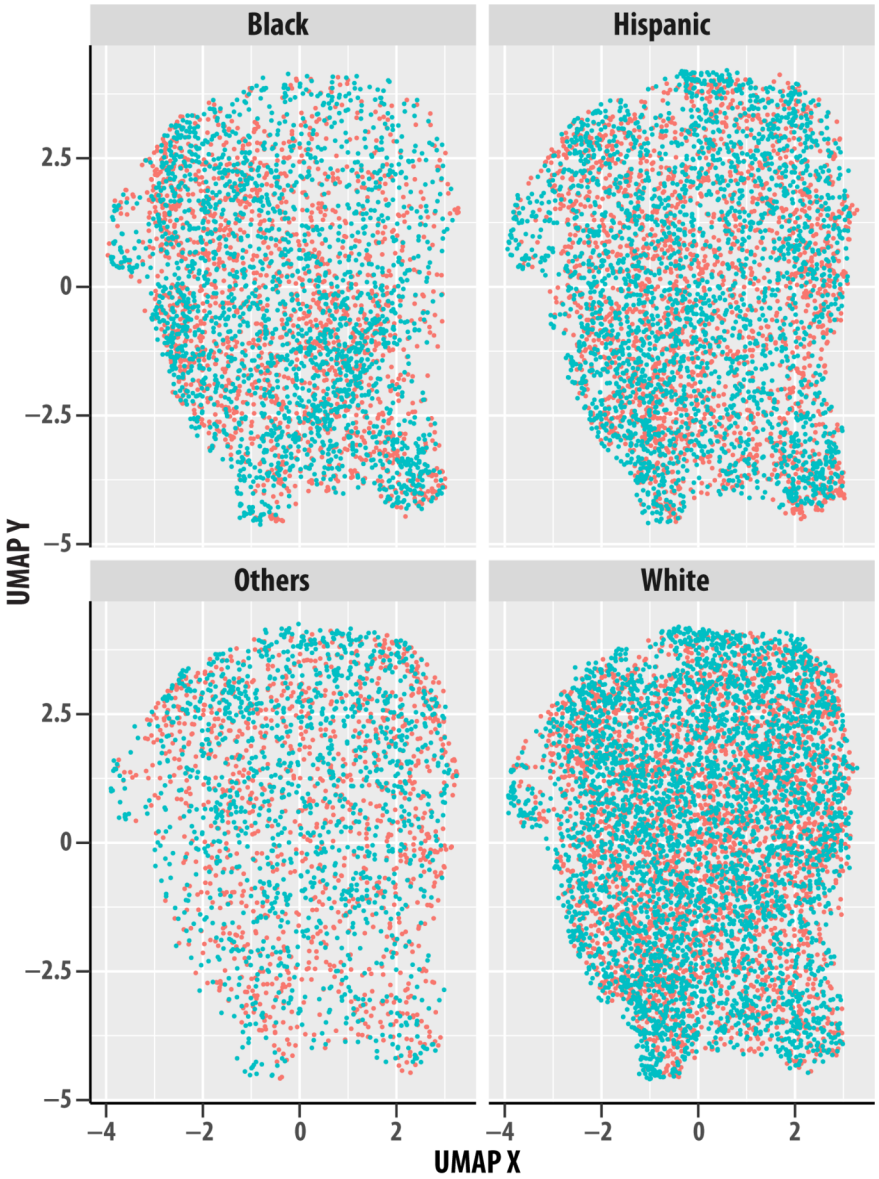
SUPPLEMENTARY FIGURE 2 (1 of 2)



SUPPLEMENTARY FIGURE 2 (2 of 2)



SUPPLEMENTARY FIGURE 3



SUPPLEMENTARY FIGURE 4