Journal of Pharmacokinetics and Pharmacodynamics (2023) 50:111–122
https://doi.org/10.1007/s10928-022-09838-4

ORIGINAL PAPER

# Generative models for age, race/ethnicity, and disease state dependence of physiological determinants of drug dosing

Rahul Nair[1] · Deen Dayal Mohan[1] · Srirangaraj Setlur[1] · Venugopal Govindaraju[1] · Murali Ramanathan[2]

Received: 7 November 2022 / Accepted: 12 December 2022 / Published online: 24 December 2022


## Abstract

Dosing requires consideration of diverse patient-specific factors affecting drug pharmacokinetics and pharmacodynamics. The available pharmacometric methods have limited capacity for modeling the inter-relationships and patterns of variability among physiological determinants of drug dosing (PDODD). To investigate whether generative adversarial networks (GANs) can learn a generative model from real-world data that recapitulates PDODD distributions. A GAN architecture was developed for modeling a PDODD panel comprised of: age, sex, race/ethnicity, body weight, body surface area, total body fat, lean body weight, albumin concentration, glomerular filtration rate (EGFR), urine flow rate, urinary albumin-to-creatinine ratio, alanine aminotransferase to alkaline phosphatase R-value, total bilirubin, active hepatitis B infection status, active hepatitis C infection status, red blood cell, white blood cell, and platelet counts. The panel variables were derived from National Health and Nutrition Examination Survey (NHANES) data sets. The dependence of GAN-generated PDODD on age, race, and active hepatitis infections was assessed. The continuous PDODD biomarkers had diverse non-normal univariate distributions and bivariate trend patterns. The univariate distributions of PDODD biomarkers from GAN simulations satisfactorily approximated those in test data. The joint distribution of the continuous variables was visualized using three 2-dimensional projection methods; for all three methods, the points from the GAN simulation random variate vectors were well dispersed amongst the test data. The age dependence trend patterns in GAN data were similar to those in test data. The histograms for R-values and EGFR from GAN simulations overlapped extensively with test data histograms for the Hispanic, White, African American, and Other race/ethnicity groups. The GAN-simulated data also mirrored the R-values and EGFR changes in active hepatitis C and hepatitis B infection. GANs are a promising approach for simulating the age, race/ethnicity and disease state dependencies of PDODD.

**Keywords** Artificial intelligence · AI · Generative adversarial networks · Pharmacometrics

## Introduction

The composition, physicochemical properties, and labeling of drug products are carefully controlled during drug development and manufacturing to assure product quality and performance. However, optimal drug dosing decisions in real-world clinical settings require consideration of the interactions of the drug product with patient-specific characteristics that can be highly variable. Factors related to the absorption, distribution, metabolism, and elimination processes responsible for drug pharmacokinetics (PK) and pharmacodynamics (PD) are particularly important patient-specific physiological determinants of drug dosing (PDODD).

The goal of the covariate modeling step in pharmacometric model development is to identify patient-specific factors that can explain variability of the PK parameters in the structural model [1, 2]. The specific sites at which the critical interactions of the drug product with physiological processes occur often cannot be sampled or characterized in the clinical setting and their potential impact must be inferred from other biomarkers that covary with these processes. Age, sex, race/ethnicity, body weight and body

✉ Murali Ramanathan
Murali@Buffalo.Edu

[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, 355 Pharmacy Building, Buffalo, NY 14214-8033, USA

surface area are examples of patient-specific covariates that are typically assessed in pharmacometric modeling because they are easily obtained. There are however an ever-increasing number of observable surrogate markers that can be leveraged as potential sources of information regarding whole body and organ-specific function, e.g., the blood, liver, and kidney, in the clinical setting.

PDODD can exhibit complex inter-dependencies that can include both non-linear multivariate trends and patterns of variability containing pairwise correlations and higher-order associations, e.g., body weight has complicated dependencies on age, sex, race/ethnicity. Conceptually, all the underlying relationships among all the salient patient-specific characteristics in the population can be represented as a high-dimensional joint distribution that subsumes the trends, pairwise correlations, and multivariate associations among the constituent variables.

Parametric methods such as multivariate distributions, copulas, and Bayesian models can characterize different aspects of the joint distribution, but these methods require extensive user input that limits their utility to small numbers of variables and known distributions [1–3]. Approaches requiring less user input, e.g., information theoretic and machine learning (ML) algorithms such as random forests, have also been investigated for identifying and modeling the inter-dependencies among key pharmacometric covariates [4, 5].

This research investigates an innovative approach that can learn models for complex high-dimensional PDODD joint distributions and their dependence on age, race/ethnicity, and disease state from real-world "big data" and generate random variate PDODD vectors. The approach uses an emerging artificial intelligence (AI) deep learning method called generative adversarial networks (GANs), which has been used to create realistic simulations of complex patterns in images [6]. Users can employ the learned model to simulate PDODD vectors without the need to access or analyze the underlying real-world data.

## Methods

### Panel of physiological determinants of drug dosing (PDODD)

*Dataset* We used public-domain data from the National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics (NCHS) in the United States. The NHANES collects data from laboratory measurements, physical screening, and surveys; the data are made available to the public every 2 years [7].

The PDODD panel consisted of age, sex, race/ethnicity, body weight, body surface area, total body fat, lean body weight, albumin concentration, glomerular filtration rate, urine flow rate, urinary albumin-to-creatinine ratio, alanine aminotransferase to alkaline phosphatase R-value, total bilirubin, active hepatitis B infection status, active hepatitis C infection status, red blood cell, white blood cell, and platelet counts.

Data required for the PDODD panel were extracted and pooled from the 2011–2012, 2013–2014, 2015–2016, and 2017–2018 NHANES data release cycles.

*Data pre-processing* Subjects 12 years and older were included. In NHANES, subjects 80 years-old and over are coded as 80 years.

Race was recoded from the *RIDRETH1* Race/Hispanic origin variable. The Mexican American and Other Hispanic participants were recoded as Hispanic; the other races (Non-Hispanic White, Non-Hispanic Black, Other – including Multiracial) were retained unchanged.

Body surface area ($BSA$, $m^2$) was calculated from $Weight(kg)$ and $Height(cm)$ using the Dubois and Dubois Eq. (8):

$$BSA = 0.007184 Weight^{0.425} Height^{0.725}$$

Estimated glomerular filtration rate ($EGFR$, ml/min 1.73 $m^2$) was obtained from serum creatinine measurements using the CKI-EPI study 2021 formula [9].

$$EGFR = 142\beta\left(0.9938^{Age}\right)\min(LBXSCR/\kappa, 1)^{\alpha}\max(LBXSCR/\kappa, 1)^{-1.2}$$

In the equation: $LBXSCR$ is serum creatinine in mg/dl; $\beta$ is a constant that is 1.012 for females and 1 for males, $Age$ is in years; $\kappa$ is a constant that is 0.7 for females and 0.9 for males, $\alpha$ is a constant that is $-$ 0.241 for females and $-$ 0.302 for males; min and max are minimum and maximum functions [9].

$RVALUE$ is a computed measure of liver function [10] obtained from serum alanine aminotransferase ($LBXSATSI$) and serum alkaline phosphatase ($LBXSAPSI$) activity measurements in a standard complete metabolic panel (CMP).

$$RVALUE = (LBXSATSI/ULN_{LBXSATSI})/(LBXSAPSI/ULN_{LBXSAPSI})$$

$ULN_{LBXSATSI}$ and $ULN_{LBXSAPSI}$ are the upper limits of normal (ULN) for alanine aminotransferase (ALT) and alkaline phosphatase (ALP), respectively. $ULN_{LBXSATSI}$ was set to 29 IU/L for males and 22 IU/L for females [11]. Based on Gonzalez et al. [12], $ULN_{LBXSAPSI}$ for the Hispanic group was 123.2 IU/L for females and 123.8 IU/L for males; for the non-Hispanic White group it was 97.1 IU/L for females and 109.6 IU/L for males; for the non-Hispanic Black group it was 109.9 IU/L for females and 116.3 IU/L for males; the values non-Hispanic White group were used

for the Other – including multi-Racial group. The previously described recoded *Race* variable was employed.

The average urine flow rate was calculated from three separation urine collections using NHANES guidelines [13].

Active hepatitis B virus (HBV) infection status was a binary variable that was set to unity for anti-HBV core antigen antibody (anti-HBc Ab, LBXHBV) positive subjects who tested positive for HBV surface antigen (HBsAg, LBDHBG) and 2 for anti-HBc Ab tested subjects not meeting the criterion. Active hepatitis C virus (HCV) infection status was a binary variable that was set to unity for anti-HCV screening antibody (anti-HCV Ab) positive subjects who tested positive for HCV-RNA (LBXHCR) and 2 for anti-HCV Ab screening antibody subjects not meeting the criterion.

The continuous biomarker data were log-transformed, and minmax scaled to the range $[-1,1]$. There were a few zeroes in the bilirubin variable; a small positive number (0.0009), which was 10-fold lower than the lowest reported measured value was added prior to log-transformation.

Data pre-processing and variable computations were conducted with the R statistical computing platform [14].

The pooled data were randomly split into training (80%) and test (20%) data sets. Listwise exclusion was employed.

### GAN architecture

A generative adversarial network (GAN) is comprised of two neural networks called generator and discriminator that are trained competitively. During training, the generator transforms random vectors from a latent space to synthesize generated data vectors. The discriminator is a binary classifier that is trained to distinguish real data vectors from generated data vectors. Upon successful training, the joint distribution of the generated data vectors approximates the joint distribution of the real data. Supplementary Fig. 1 shows the characteristic features of a typical GAN and describes the key functions of its components.

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 [15]. A key consideration during the design of the GAN architecture for PDODD modeling was the tabular nature of the data, which lacks the local correlation structures present in image analysis [16].

The model was trained for 1000 epochs with batch size of 300 and five discriminator steps. Upon successful training, a simulated dataset containing 10,000 GAN-simulated random variate data vectors was computed. The GAN-generated data distributions were compared to the distribution of the test data.

The GANs for modeling PDODD were prototyped using PyTorch, an open-source library for AI and machine learning based on the Python programming language [17].

### Data analysis

Data analyses and visualization were conducted in the R statistical computing platform [14].

In exploratory analysis, descriptive statistics (frequencies, mean, standard deviations, median, inter-quartile range. and range) were obtained from the input data set prior to GAN modeling. The *ggpairs* package was used to generate pairs panel plots containing bivariate densities, univariate densities, and bivariate scatter plots with loess fits of the input dataset. Box plots of RVALUE and EGFR in the groups with and without active hepatitis C or active hepatitis B infection were obtained and independent sample *t*-tests were used to evaluated differences between the groups.

GAN performance was assessed by comparing the GAN-generated biomarker distributions to the test data. A random sample of GAN-generated variates of the same size as the test data set was employed.

For visualization of univariate distributions of the continuous PDODD variables, probability density histograms and quantile-quantile plots (QQ plots) of test data vs. GAN-generated data were compared.

For visualization of the multivariate distribution, 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 [18–20]. The perplexity and theta hyperparameters for t-SNE were set to 50 and 0.5, respectively.

## Results

### Study population and biomarker panel

The study data were a subset from the population based NHANES study participants ($n$ = 27,832). We excluded participants ($n$ = 11,324) who were less than 12-years-old

age at the screening visit. Table 1 summarizes the demographic characteristics and descriptive statistics for the drug disposition biomarkers for $n = 27{,}832$ participants included.

Supplementary Fig. 2 summarizes the probability mass functions of the discrete PDODD variables: sex, race/ethnicity; active hepatitis B virus and hepatitis C virus infection status are also summarized. The proportions of males and females was similar across the different race/ethnicity groups. The overall of active hepatitis B and hepatitis C infection frequencies were 0.59% and 1.24% (Table 1), respectively.

The pairs plots in Fig. 1A and B summarize the univariate densities (along diagonal), bivariate densities (contour density plots in lower triangular region) and trends (loess fits in bivariate plots in upper triangular region) among the log transformed and minmax scaled continuous PDODD variables. The set of variables was divided into two sets of seven variables so that the number of plots in the pair plot panel was manageable. The loess fits in Fig. 1A and B highlight the different non-linear trends among the continuous variables. The bivariate contour density plots show that a variety of different non-normal distribution patterns of variations are present among the PDODD.

## Continuous PDODD panel joint distribution simulations

We compared the random-variate data vectors from GAN simulations to the test data. The multivariate joint distribution of the PDODD panel contains 14 continuous variables and four categorical variables (sex, race/ethnicity, active hepatis C infection and hepatis B infection) cannot be visualized directly.

We first evaluated GAN for simulating the continuous variables in the PDODD panel containing 14 continuous variables using univariate probability density histograms and quantile-quantile plots (QQ plots) of the empirical data distributions. The probability density histograms of GAN-simulated data are compared to the corresponding test data in Fig. 2 A-N for the 14 continuous variables in the PDODD panel. There was extensive overlap of the GAN-simulated univariate distributions with the test data univariate distributions. The QQ plots, (Supplementary Fig. 2) were all clustered around the line of identity, which confirms satisfactory approximation of the univariate distributions of the continuous variables by the GAN.

In the next step, we assessed 2-dimensional projections of the joint distribution of the 14 continuous variables. We used three separate projection methods: t-SNE, UMAP and PCA. The results in Fig. 3 show that the projections of the GAN-simulated data vectors were well dispersed among the projections of the test data vectors for all three methods. The loess lines, which were used to evaluate regions of deviation, showed only modest deviations. These evaluations are consistent with a satisfactory approximation of the multivariate joint distribution of the PDODD panel.

Many PDODD exhibit complex patterns of age dependence. To further assess GAN approximation of the multivariate joint distribution of the PDODD panel, we visualized the age-dependence of bivariate distributions using scatter plots (Fig. 4). We used loess lines to determine whether the inter-dependence trends of the continuous PDODD variables from the GAN simulation were satisfactory approximated the trends in test data. The GAN-simulated data points were well dispersed with the test data points in bivariate scatter plots with age. The loess lines overlapped extensively. This indicates that GAN simulations provide a satisfactory model for PDODD age dependence trends and variability patterns.

## Conditional GAN simulations of race/ethnicity and disease states on PDODD

The categorical and continuous variables in the PDODD panel were modeled simultaneously in our GAN approach. We conducted assessments of the conditional distributions of RVALUE and EGFR with the categorical variables of race/ethnicity, active hepatitis C infection status and active hepatitis B infection status. We selected RVALUE and EGFR, which are biomarkers of hepatic and renal function, respectively, as representative continuous PDODD for these conditional assessments because of the importance of liver and kidney function in drug metabolism and elimination.

*Effect of race/ethnicity* The percentage of females were similar in the GAN-simulated data (51.9%) and test data (52.7%). The percentages of Hispanics and African Americans were similar in the GAN-simulated data (30.1% Hispanic and 18.4%) and test data (30.3% and 18.5%).

Figure 5 shows the probability density histograms of RVALUE and EGFR variables in the Hispanic, White, African American, and Other race/ethnicity groups. There was extensive overlap between the GAN-generated data and the test data histograms. This indicates the potential utility of the GAN approach for incorporating race/ethnicity differences in PDODD.

*Effect of disease state* Some disease states can cause alterations to PDODD. We used hepatitis B and hepatitis C as an exemplar for evaluating whether GAN models could recapitulate PDODD distributions in disease. Active hepatitis B and hepatitis C status were inferred from the laboratory test results in NHANES. Active hepatitis C and

**Table 1** Detailed summary statistics of the demographics and biomarkers from data set combined from NHANES 2011–2012, 2013–2014, 2015–2016 and 2017–2018

| Variable name | Variable description | Actual N (% Missing) | Percent | | |
|---|---|---|---|---|---|
| RIAGENDR | Sex | 27,832 (0) | – | | |
| | Female | 14,258 (0) | 51.2 | | |
| | Male | 13,574 (0) | 48.8 | | |
| RIDETH1 | Race/Ethnicity | 27,832 (0) | – | | |
| | Non-Hispanic Black | 6422 | 23.1% | | |
| | Mexican American | 4095 | 14.7% | | |
| | Other Hispanic | 2915 | 10.5% | | |
| | Non-Hispanic White | 9678 | 34.8% | | |
| | Other–Including multi-racial | 4722 | 17.0% | | |
| HEPB | Active hepatitis B infection status | 145/24,767 (11%) | 0.59% | | |
| HEPC | Active hepatitis C infection status | 232/18,422 (33%) | 1.24% | | |
| | | **N missing (%)** | **Mean (SD)** | **Median (IQR)** | **Min – Max** |
| RIDAGEYR | Age in years at screening | 0 (0) | 43.3 (20.9) | 42 (24–61) | 12–80 |
| BMXWT | Weight (kg) | 1495 (5.4) | 78.8 (22.5) | 75.7 (62.9–90.7) | 27.7–243 |
| BSA | Body surface area ($m^2$) | 1534 (5.5) | 1.86 (0.268) | 1.84 (1.67–2.03) | 1.02–3.1 |
| DXDTOFAT | Total fat (g) | 12,789 (46) | 25,500 (12,000) | 23,400 (16,800–31,800) | 4900- 102,000 |
| DXDTOLE | Total lean excluding bone mineral content (g) | 12,499 (44.9) | 49,700 (12,900) | 48,278 (39,700–58,100) | 19,800–112,000 |
| LBXSAL | Albumin, serum (g/dL) | 3142 (11.3) | 4.25 (0.361) | 4.3 (4–4.5) | 2- 5.6 |
| URDACT | Albumin creatinine ratio (mg/g) | 1812 (6.5) | 48.7 (663) | 7.41 (4.78–14.4) | 0.21–92,500 |
| LBXRBCSI | Red blood cell count (million cells/uL) | 2657 (9.5) | 4.69 (0.503) | 4.68 (4.36–5.02) | 1.67–8.3 |
| LBXWBCSI | White blood cell count (1000 cells/uL) | 2657 (9.5) | 7.21 (3.44) | 6.9 (5.7–8.4) | 1.4–400 |
| LBXPLTSI | Platelet count (1000 cells/uL) | 2658 (9.6) | 241 (61.3) | 235 (200–276) | 8- 818 |
| URDFLOW | Urine flow rate average (mL/min) | 4007 (14.4) | 1.1 (1.36) | 0.783 (0.481–1.34) | 0.006–76.7 |
| EGFR | Glomerular filtration rate indexed ml/(min 1.73 $m^2$) | 3145 (11.3) | 101 (26) | 102 (84.8–120) | 2.02–194 |
| RVALUE | Ratio of alanine aminotransferase to alkaline phosphate | 3152 (11.3) | 1.46 (1.18) | 1.23 (0.846–1.75) | 0.0509- 43.3 |
| LBXSTB | Total bilirubin (mg/dL) | 3157 (11.3) | 0.591 (0.308) | 0.5 (0.4–0.7) | 0- 7.1 |

The total number of cases was $n$ = 29,547 cases

RIDAGEYR is age in years and subjects 80 years-old and over are coded as 80 years. Only subjects 12 years and older included.BSA was calculated from *Weight* (*kg*) and *Height* (*cm*) using the Dubois and Dubois equation:$BSA(m^2) = 0.007184 Weight(kg)^{0.425} Height(cm)^{0.725}$. RVALUE is a measure of hepatic injury based on the ratio of serum alanine aminotransferase and alkaline phosphate activities normalized to their respective upper limits of normal. EGFR is glomerular filtration rate calculated from serum creatinine and age in ml/min per 1.73 $m^2$ of body surface area. URDFLOW is a derived variable calculated from 3 possible urine flow measurements. HEPB is based on positivity for both HBV surface antigen and anti-HBVc Ab. HEPC is based on positivity for both HCV-RNA and anti-HCV Ab

hepatitis B infection were infrequent (1.24% and 0.59%, respectively; Table 1).

Supplementary Fig. 4 shows box plots of RVALUE and EGFR in the groups with and without active hepatitis C or active hepatitis B infections in the NHANES-derived data prior to GAN modeling. The RVALUES were higher in the group with active hepatitis C (mean: 3.52 vs.1.42, $p < 0.001$, $t$-test) and in the group with active hepatitis B (1.87 vs. 1.46, $p = 0.001$, $t$-test); the EGFR was lower in the group with active hepatitis C (87.7 vs. 101 ml/min 1.73 $m^2$, $p < 0.001$) and modestly lower in the group with active hepatitis B (96.5 vs. 101 ml/min 1.73 $m^2$, $p = 0.019$).

There were 19 subjects and 1863 subjects without active hepatitis C in the GAN-simulated data (1.00%) compared to 15 subjects with active hepatitis C and 1867 subjects

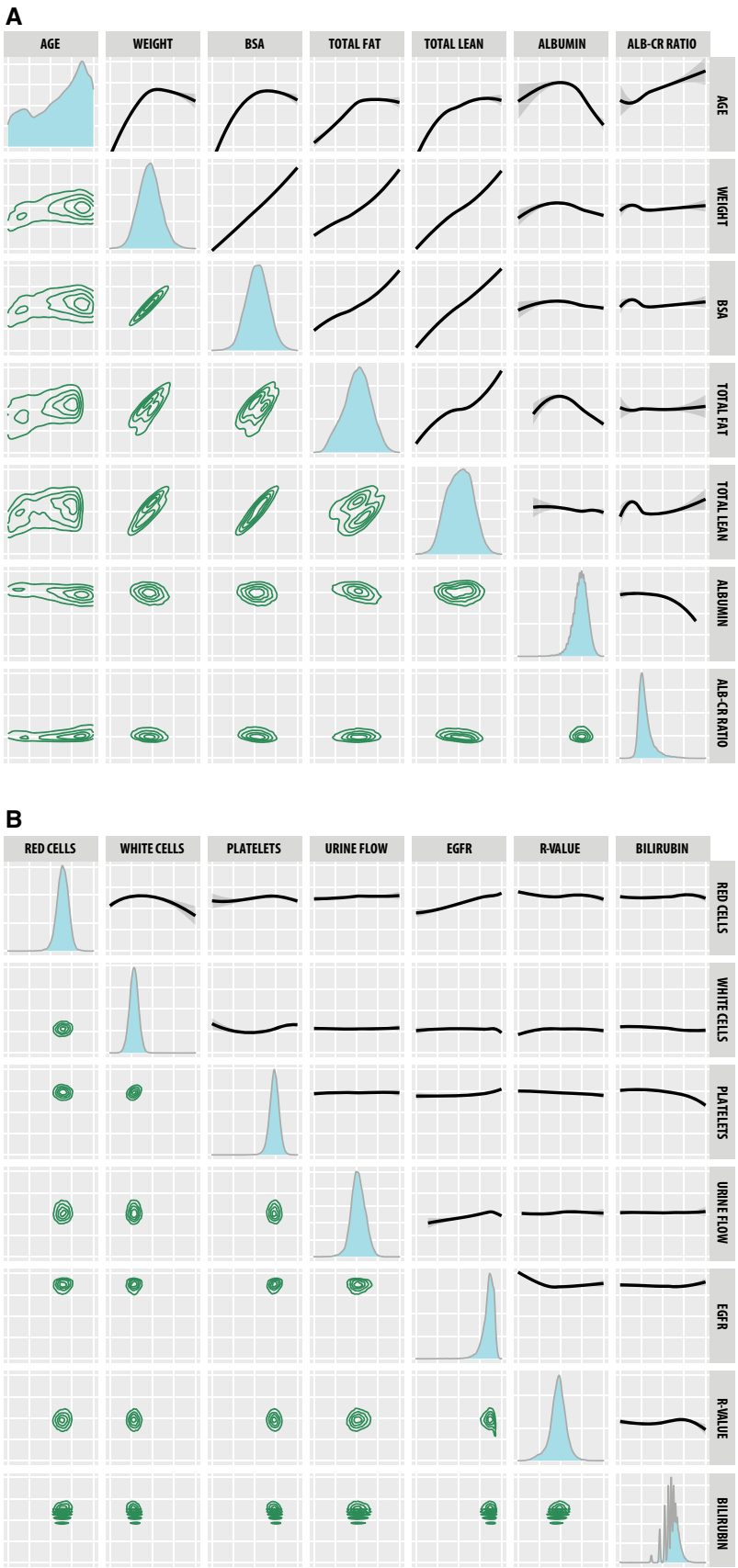


**Fig. 1** **A** is a pairs panel plot of seven continuous variables: age, weight, body surface area (BSA), total fat mass, total lean mass, albumin, and urine albumin-to-creatinine ratio. **B** is the corresponding pairs panel plot of seven other continuous variables: red blood cell count, white blood cell count, platelet count, urine flow, estimated glomerular filtration rate (EGFR), alanine aminotransferase to alkaline phosphatase R-value, and bilirubin. All variables were log transformed and minmax scaled to the range [− 1, 1]. The diagonal contains the univariate probability density functions. The lower triangular region represents the bivariate density of the variables along the row and column as a contour plot (green lines). The upper triangular region shows loess fit lines (black lines) to the bivariate scatter plots (points are not shown to reduce clutter) of the variables along the row and column. The light gray shadows around the loess fit lines are confidence intervals of the fit

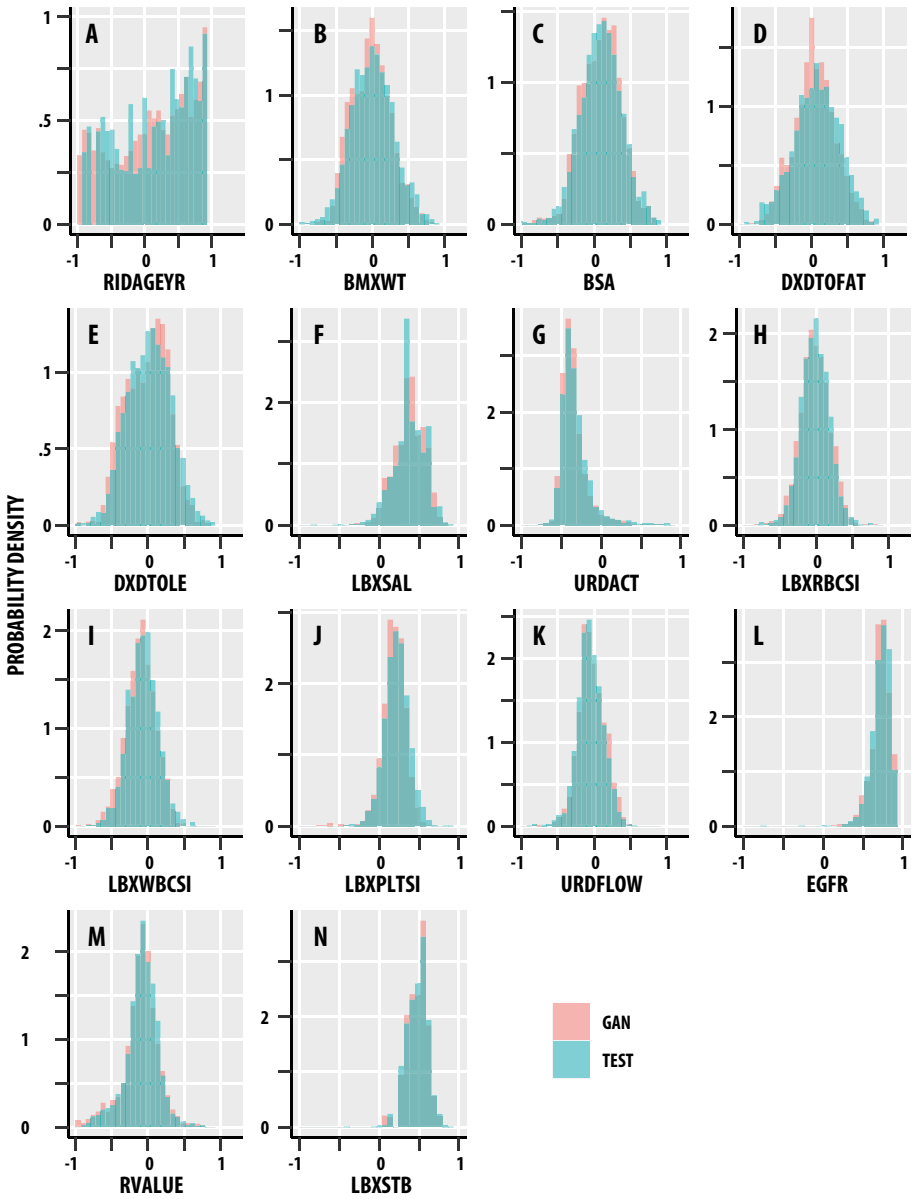


**Fig. 2** It compares the univariate probability density histograms from test data (teal bars) to the corresponding GAN-simulated values. The overlap between the two histograms is shown in the darker gray-teal. The *x*-axes of Fig. 2 A-N, respectively, correspond to logarithm and minmax transformed values of RIDAGEYR: Age in years at screening; BMXWT: Weight (kg); BSA: Body surface area ($m^2$); DXDTOFAT: Total fat (g); DXDTOLE: Total lean excluding bone mineral content (g); LBXSAL: Albumin, serum (g/dL); URDACT: Albumin-creatinine ratio (mg/g); LBXRBCSI: Red blood cell count (million cells/μL); LBXWBCSI: White blood cell count (1000 cells/μL); LBXPLTSI: Platelet count (1000 cells/μL); URDFLOW: Urine flow rate average (mL/min); EGFR: Glomerular filtration rate indexed ml/(min 1.73 $m^2$); RVALUE: Ratio of alanine aminotransferase to alkaline phosphate; LBXSTB: Total bilirubin (mg/dL)

without active hepatitis C in the test data (0.80%); the frequencies differences were not different ($p = 0.61$, Fisher exact test). The frequencies of active hepatitis B were also not different between the GAN-simulated and test data sets ($p = 0.84$, Fisher exact test).

Figure 6 shows the probability density histograms of the GAN-simulated RVALUE, EGFR values in the groups with active hepatitis C status or active hepatitis B. Again, there was extensive overlap of the GAN histograms with the test data histograms in all the groups.

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

## Discussion

In this research we developed and evaluate an innovative approach for generative modeling of the joint distribution of PDODD and the dependencies on age, race/ethnicity, and disease state.

In addition to demographic characteristics such as age, sex, and race/ethnicity, we included a panel of diverse biomarkers that are important determinants of dosing decisions, drug disposition, and treatment outcomes across therapeutic classes. Body weight and body surface area are widely used for individualizing doses clinically. We also included total fat and lean body mass measurements (from

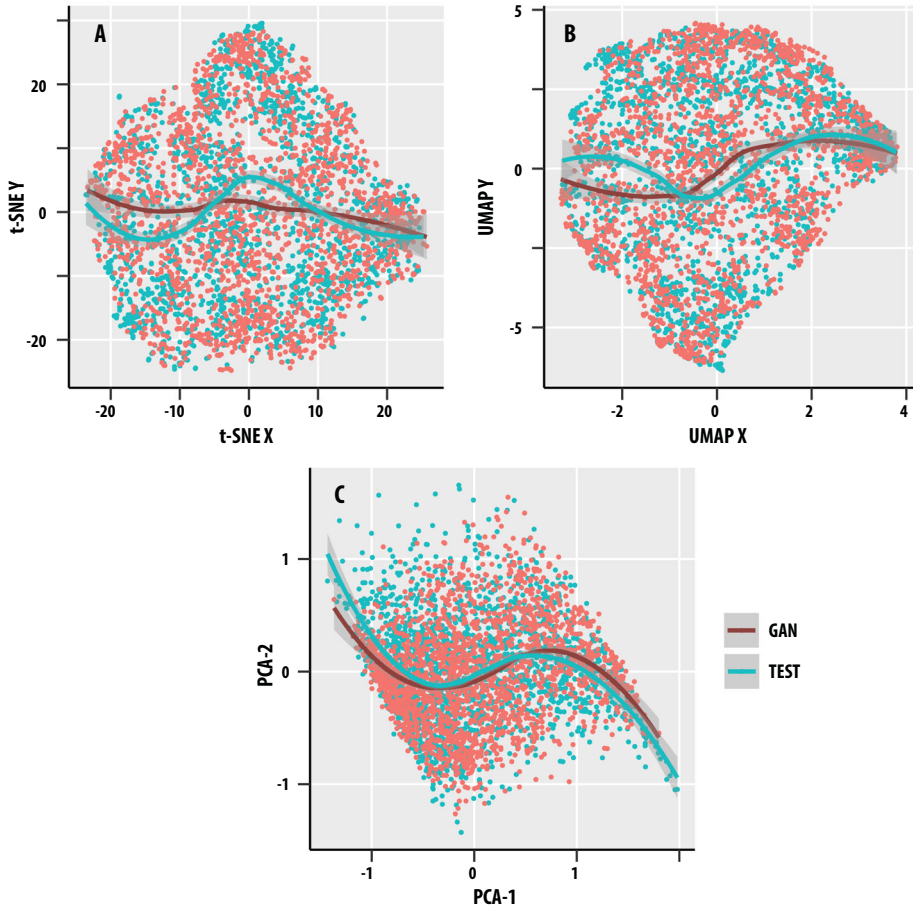


**Fig. 3** The scatter plots show the 2-dimensional *t*-stochastic neighbor embedding (t-SNE, **A**), uniform manifold approximation and projection (UMAP, **B**) and principal components analysis (PCA, **C**) projections of the continuous variables from the test data (teal points) and the GAN-simulated data (salmon points). The solid lines represent the loess fits to the test data (teal line) and the GAN-simulated data (dark red line); the gray envelope represents the confidence interval around the loess lines. The 2-dimensional projections were obtained for logarithm and minmax transformed values of RIDAGEYR: Age in years at screening; BMXWT: Weight (kg); BSA: Body surface area ($m^2$); DXDTOFAT: Total fat (g); DXDTOLE: Total lean excluding bone mineral content (g); LBXSAL: Albumin, serum (g/dL); URDACT: Albumin-creatinine ratio (mg/g); LBXRBCSI: Red blood cell count (million cells/μL); LBXWBCSI: White blood cell count (1000 cells/μL); LBXPLTSI: Platelet count (1000 cells/μL); URDFLOW: Urine flow rate average (mL/min); EGFR: Glomerular filtration rate indexed ml/(min 1.73 $m^2$); RVALUE: Ratio of alanine aminotransferase to alkaline phosphate; LBXSTB: Total bilirubin (mg/dL) (Color figure online)

dual energy X-ray absorptiometry) as hydrophobic drugs preferentially partition into adipose tissue and lean body mass is useful for dosing in obesity [21]. Albumin is the most abundant plasma protein and albumin binding is important for many acidic drugs [22, 23]. Some drugs partition extensively into red blood cells causing discrepancies between blood and plasma drug concentrations [24]. We included white blood cell and platelet counts (obtained from the complete blood count) because baseline levels of these cells can affect dosing decisions for drugs that cause lymphopenia, neutropenia, and thrombocytopenia. We used several measures of hepatic and renal function because the liver and kidney are important sites for drug metabolism, transport, and elimination. The renal function biomarkers included urine flow rate, glomerular filtration rate, and urine albumin to creatinine ratio. The albumin-to-creatine ratio assesses proteinuria and could be a predictor for high renal clearance of protein drugs. The hepatic biomarkers included the alanine aminotransferase to alkaline phosphatase ratio RVALUE, which is useful for distinguishing hepatocellular liver injury from cholecystic disease and bilirubin. Hepatocellular injury with jaundice forms the basis for Hy's law, which is a reliable approach for predicting drug-induced liver injury [25, 26].

Our approach, which included both whole body and organ-specific measures, has several distinctive features useful for drug disposition modeling but also weaknesses, many of which might be addressable. Our PDODD panel could be criticized for lacking biomarkers for specific drug classes or diseases. Because the primary purpose here was

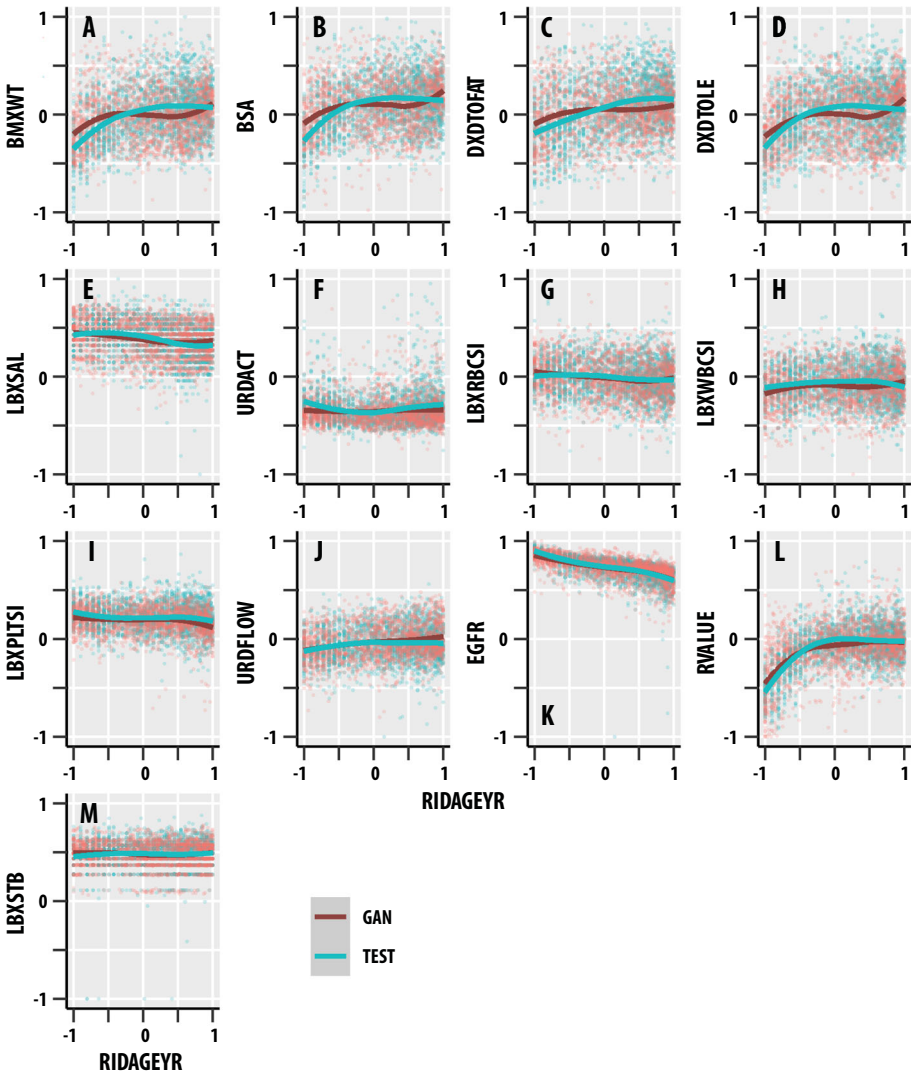


**Fig. 4** The scatter plots show the age dependence of the test data (teal points) and the GAN-simulated data (salmon points). The solid lines represent the loess fits to the test data (teal line) and the GAN-simulated data (dark red line). The *x*-axis on all figures is logarithm and minmax transformed values of RIDAGEYR: Age in years at screening; The *y*-axes of Fig. 4 A-M, respectively, correspond to logarithm and minmax transformed values of BMXWT: Weight (kg); BSA: Body surface area ($m^2$); DXDTOFAT: Total fat (g); DXDTOLE: Total lean excluding bone mineral content (g); LBXSAL: Albumin, serum (g/dL); URDACT: Albumin-creatinine ratio (mg/g); LBXRBCSI: Red blood cell count (million cells/μL); LBXWBCSI: White blood cell count (1000 cells/μL); LBXPLTSI: Platelet count (1000 cells/μL); URDFLOW: Urine flow rate average (mL/min); EGFR: Glomerular filtration rate indexed mL/(min 1.73 $m^2$); RVALUE: Ratio of alanine aminotransferase to alkaline phosphate; LBXSTB: Total bilirubin (mg/dL) (Color figure online)

to conduct rigorous proof-of-concept for problems relevant to drug disposition, we intentionally selected measures that were easily and reliably obtained in the routine clinical setting. However, not all measures routinely available in a test were included. In certain cases, we were limited by the date collected in NHANES, e.g., albumin was included but we did not have data for alpha-1 acid glycoprotein or oromucosoid, which binds basic and neutral drugs [22, 23]. Among the disease and disease exposures for which we had data available, we included hepatitis because it can cause liver damage and because positivity for hepatitis is relatively common.

Notably our PDODD panel consisted of two classes of biomarkers. One class of PDODD biomarkers was directly extracted from clinical test results (e.g., weight, urine flow, cell counts, albumin, bilirubin) whereas PDODD biomarkers in the other class were computed from two or more direct measurements, e.g., body surface area, EGFR and RVALUE. We did not consider large scale "omics" data as the clinical utilization of these methods is not established. Based on the proven versatility of GAN for modeling complex, high dimensional distributions in image generation applications, we anticipate that larger panels of biomarkers could also be accommodated.

GAN provide an elegant approach for the statistical problem of modeling high dimensional joint distributions, which is particularly challenging because the data contains higher-order correlation patterns. The GAN approach is a generative model because it can be trained to approximate high-dimensional joint distributions from the instances presented from the datasets. GAN do not require the user to specify the functional form of the joint distribution; the distribution structure is encoded in the neural networks during training. Once successfully trained, the generator can be used to generate random data vector variates concordant with the underlying data

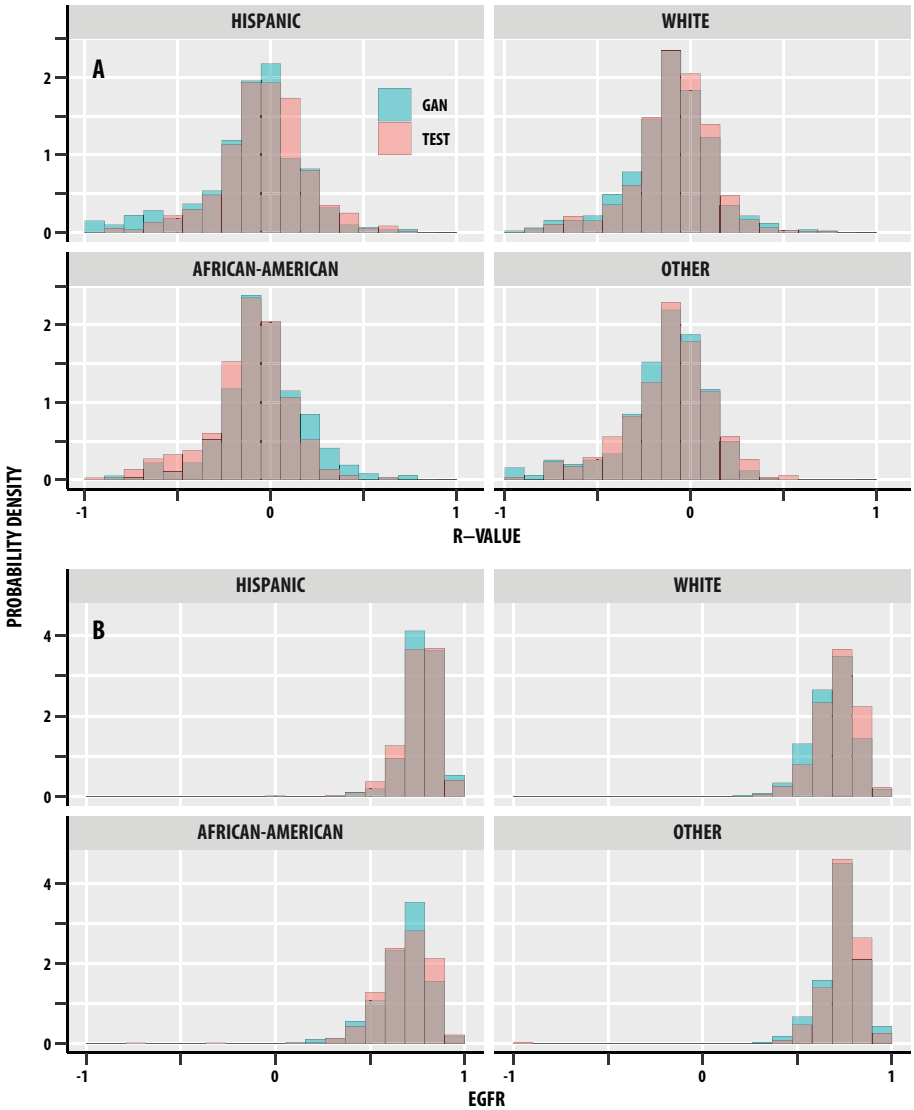


**Fig. 5** It compares the probability density histograms of the GAN-simulated values (teal bars) of R-value (Fig. 5 A) and EGFR (Fig. 5B) to the test data (salmon bars) in the Hispanic, White, African American, and Other race/ethnicity groups. The region of overlap between the histogram bars is in the darker brown shade

distribution. Examples of alternatives to the generative modeling are resampling, Bayesian modeling, and copula-based methods. Resampling methods provide sample vector sets that are verbatim subsamples of the data set used. Bayesian modeling requires priors, which can be challenging to specify for high-dimensional data sets. However, Bayesian models are useful in the setting of modeling the distributions of multivariate functions of random variables. The mathematical foundations of copulas are more complex than resampling and Bayesian models. Copula methods require the user to specify the marginal distributions of each individual biomarker and additionally provide a copula distribution to encode functional form for the higher-order correlation structure amongst the variables.

Our results demonstrate the utility and elucidate the potential of the GAN approach for modeling PDODD.

## Study highlights

**What is the current knowledge on the topic?** Personalized medicine requires consideration of patient-specific factors that affect drug disposition. Age, race, hepatic and renal function disease states, and other factors can affect the dosing decisions.

**What question this study addressed?** To evaluate generative adversarial networks (GANs), a deep learning-based artificial intelligence technology, for modeling patient specific physiological determinants of drug dosing.

**What this study adds to our knowledge?** The results demonstrate that GANs can be used to generate simulated samples that mimic the joint distribution of complex physiological determinants of drug dosing.

**How this might change clinical pharmacology and therapeutics?** The GAN approach could be a powerful and

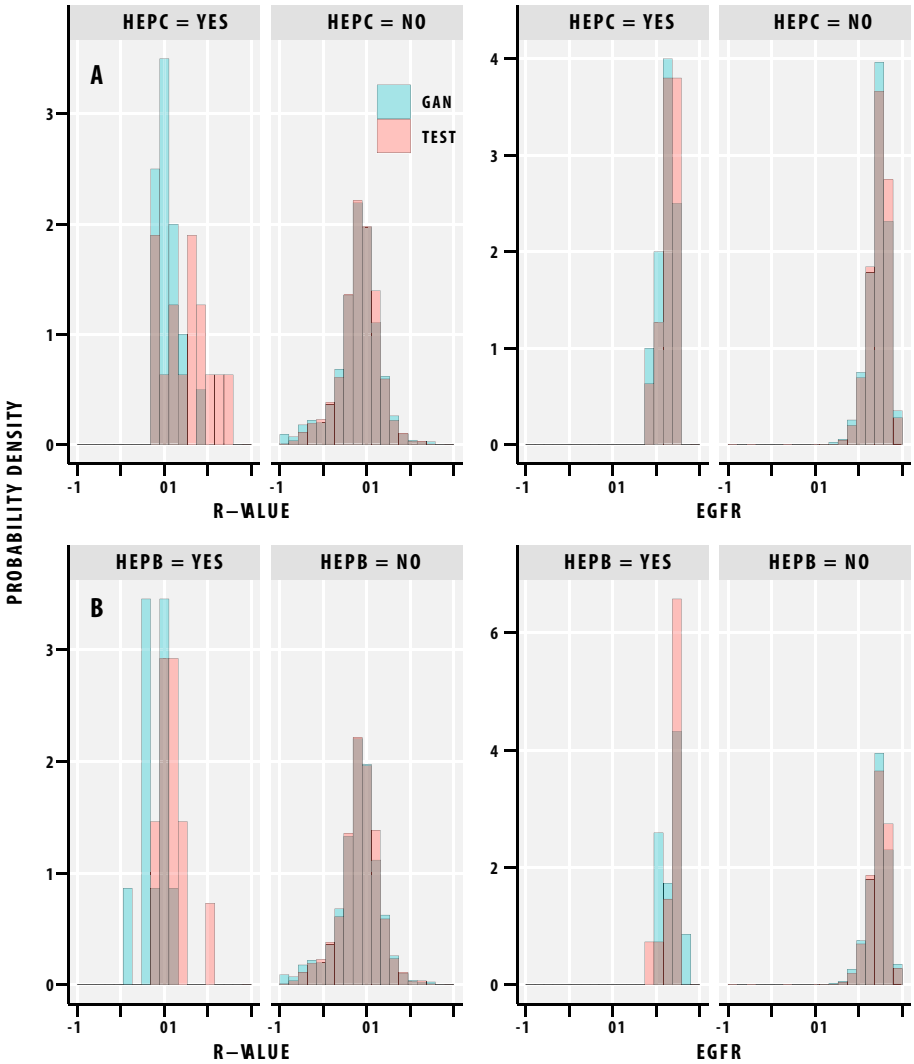


**Fig. 6** It compares the probability density histograms of the GAN-simulated values (teal bars) of R-value (left column) and EGFR (right column) to the test data (salmon bars) in the groups with (HEPC = YES) and without (HEPC = NO) active hepatitis C infection (top row) and in the group with (HEPB = YES) and without (HEPB = NO) active hepatitis B infection (bottom row). The region of overlap between the histogram bars is in the darker brown shade

versatile method for generating disease-relevant biomarker profiles and virtual patient datasets for clinical trial simulations and pharmacometrics.

**Author contributions** RN–Conducted experiments, data analysis, manuscript preparation. DDM–Assisted with experiments, data analysis, manuscript preparation. SS–Study concept and design, data analysis, manuscript preparation. VG–Study oversight, manuscript review. MR–Study concept and design, data analysis, manuscript preparation.

## Declarations

**Conflict of interest** Rahul Nair and Deen Dayal Mohan have no conflicts. Srirangaraj Setlur and Dr. Venu Govindaraju received unrelated research funding from the National Science Foundation, United States Postal Service, and the Intelligence Advanced Research Projects Activity agencies. Dr. Murali Ramanathan received research funding from the National Science Foundation, Department of Defense, and the National Institutes of Health.

**Supplementary Information**

The online version contains supplementary material available at https://doi.org/10.1007/s10928-022-09838-4.

## References

1. McComb M, Bies R, Ramanathan M (2022) Machine learning in pharmacometrics: opportunities and challenges. Br J Clin Pharmacol 88:1482–1499
2. McComb M, Ramanathan M (2020) Generalized pharmacometric modeling, a Novel paradigm for integrating machine learning algorithms: a case study of metabolomic biomarkers. Clin Pharmacol Ther 107:1343–1351
3. Knights J, Heidary Z, Peters-Strickland T, Ramanathan M (2019) Evaluating digital medicine ingestion data from seriously mentally ill patients with a bayesian hybrid model. NPJ Digit Med 2:20
4. Knights J, Sato Y, Kaniwa N, Saito Y, Ueno H, Ramanathan M (2014) Genetic factors associated with gemcitabine pharmacokinetics, disposition, and toxicity. Pharmacogenet Genomics 24:15–25

5. McComb M, Blair RH, Lysy M, Ramanathan M (2022) Machine learning-guided, big data-enabled, biomarker-based systems pharmacology: modeling the stochasticity of natural history and disease progression. J Pharmacokinet Pharmacodyn 49:65–79
6. Goodfellow IJ et al (2014) Generative Adversarial Networks. *arXiv*, arXiv:1406.2661 [stat.ML]
7. National Health and Nutrition Examination Survey (2017) About the National Health and Nutrition Examination Survey. National Center for Health Statistics, Hyattsville, MD
8. Dubois D, Dubois EF (1916) A formula to estimate the approximate surface area if height and weight be known. Arch Intern Med 17:863–871
9. Inker LA et al (2021) New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med 385:1737–1749
10. Chalasani NP et al (2014) ACG clinical guideline: the diagnosis and management of idiosyncratic drug-induced liver injury. *Am J Gastroenterol* **109**, 950–966; quiz 67
11. Ruhl CE, Everhart JE (2012) Upper limits of normal for alanine aminotransferase activity in the United States population. *Hepatology* **55**, 447–454
12. Gonzalez H et al (2020) Normal alkaline phosphatase levels are dependent on race/ethnicity: NationalGEP Health and Nutrition Examination Survey data.BMJ Open Gastroenterol**7**, 1
13. National Health and Nutrition Examination Survey (2015) National Health and Nutrition Examination Survey: NHANES 2015–2016 overview. (ed. National Center for Health Statistics) (Centers for Disease Control)
14. R Core Team. R: a language and environment for statistical computing (2022)
15. Lin Z, Khetan A, Fanti G, Oh S(2017) PacGAN: the power of two samples in generative adversarial networks. *arXiv*, arXiv:1712.04086
16. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K Modeling tabular data using conditional GAN In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) eds. Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R.)
17. Paszke A et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32 (NIPS*, **32**, (2019)
18. Krijthe JH, Rtsne (2015) T-distributed stochastic neighbor embedding using Barnes-Hut implementation
19. van der Maaten LJP, Hinton GE (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579–2605
20. McInnes L, Healy J, Melville JUMAP (2018) Uniform manifold approximation and projection for dimension reduction. *arXiv:180203426 [statML]*
21. Barras M, Legg A (2017) Drug dosing in obese adults. Aust Prescr 40:189–193
22. Bteich M (2019) An overview of albumin and alpha-1-acid glycoprotein main characteristics: highlighting the roles of amino acids in binding kinetics and molecular interactions. Heliyon 5:e02879
23. Parikh HH et al (2000) A rapid spectrofluorimetric technique for determining drug-serum protein binding suitable for high-throughput screening. Pharm Res 17:632–637
24. Hinderling PH (1997) Red blood cells: a neglected compartment in pharmacokinetics and pharmacodynamics. *Pharmacol Rev* **49**, 279–295
25. Center for Biologics Evaluation and Research & Center for Drug Evaluation and Research. Guidance for Industry: Drug-Induced Liver Injury: Premarketing Clinical Evaluation (2009)
26. Zimmerman HJ (1978) Drug-induced liver disease. Drugs 16:25–45