# Privacy-Preserving Data Synthesis via Differentially Private Normalizing Flows with Application to Electronic Health Records Data

**Bingyue Su*, Yu Wang*, Daniele Schiavazzi, Fang Liu†**

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame
Notre Dame, IN 46530 USA
{bsu1, ywang50, dschiavazzi, fliu2}@nd.edu

## Abstract

Medical data often contain sensitive personal information about individuals, posing significant limitations to it being shared or released for downstream learning and inferential tasks. We use normalizing flows (NF), a family of deep generative models, to estimate the probability density of a dataset with differential privacy (DP) guarantees, from which privacy-preserving synthetic data are generated and released. We apply the technique to an electronic health records dataset containing patients with pulmonary hypertension. We assess the learning and inferential utility of synthetic data by comparing the accuracy of hypertension predictions and the variational posterior distribution of the parameters in a physics-based model. The results suggest that synthetic data generated via NF with DP can yield good utility at a reasonable privacy cost. Our study provides evidence and adds to the growing literature on the feasibility of generating synthetic medical data for sharing or obtaining inferences from medical data using deep generative models with formal privacy guarantees.

## Introduction

Medical information is highly sensitive and may reveal patients' medical conditions and history, among others. When sharing medical data among researchers or releasing information to the public, there are always privacy concerns and risks for re-identification or disclosure of sensitive information. Even with key identifiers removed, an adversary may still be able to identify an individual or learn his or her sensitive information, leveraging publicly available data and sophisticated attacks. To mitigate privacy concerns, various privacy concepts have been proposed over the past two decades. We focus on differential privacy (DP) (Dwork et al. 2006b,a), a popular privacy framework in contemporary privacy research. Besides providing mathematical guarantees on privacy and being robust to various privacy attacks, DP has attractive properties, such as privacy loss composition and immunity to post-processing, which have contributed to its popularity for both research and practical applications.

Differentially private data synthesis (DIPS) provides a solution to integrate formal privacy guarantees into data synthesis (Rubin 1993), a common statistical disclosure limitation technique. Interested readers can refer to (Bowen and Liu 2020) for an overview of some recent DIPS techniques. Data synthesis is actively explored for sharing medical data in academia and practice (the SHARED team 2023; Dash et al. 2019; Rankin et al. 2020; Yale et al. 2020; Benaim et al. 2020; Chen et al. 2021). The introduction of variational autoencoders (Kingma and Welling 2013), generative adversarial networks (Goodfellow et al. 2014), and normalizing flow (NF) (Rezende and Mohamed 2015; Papamakarios, Pavlakou, and Murray 2017), opened new possibilities for data synthesis through deep-neural-networks-based generative models with DP guarantees (Xie et al. 2018; Jordon, Yoon, and Van Der Schaar 2018; Chen et al. 2018; Beaulieu-Jones et al. 2019; Pfitzner and Arnrich 2022).

In this work, we use electronic health records (EHR) as an example to present approaches for privacy-preserving medical data synthesis and analysis. EHR data often contain sensitive medical information about individual patients. Due to privacy concerns, it is often problematic to access single-center EHR datasets and even more challenging to create large multi-center datasets for research purposes. As a possible remedy, we examine differentially private NF (DP-NF) to generate synthetic data via density estimation with DP guarantees under a pre-defined privacy budget. To our knowledge, only a couple of works exist on DP-NF for density estimation or synthetic data generation. Waites and Cummings (2021) used the moment account method (Abadi et al. 2016) to track privacy loss in the $(\epsilon, \delta)$-DP framework during the DP-NF optimization (Waites and Cummings 2021) and applied estimated privacy-preserving densities to anomaly detection. Lee et al. (2022) studied DP-NF in the Rényi DP framework (Mironov 2017), evaluating the marginal distribution and correlations among attributes in tabular data and classification tasks in large-scaled benchmark datasets.

In contrast to the above existing work, we apply Gaussian DP (Dong, Roth, and Su 2022) to achieve privacy guarantees. Gaussian DP yields a tighter bound for the privacy loss composition and is shown to yield higher prediction accuracy in deep learning (Bu et al. 2020), compared to the moment account technique. In addition, we examine the utility of synthetic data through downstream learning and inferential tasks; the latter has not been explored in the literature for

*These authors contributed equally.
†Corresponding author.

synthetic data generated through DP-NF to our knowledge. Finally, the EHR dataset examined in this work has a small sample size, a relatively large number of continuous or discrete numerical attributes, and missing values (the training data contains 62 patients on 19 attributes), a more challenging problem for generating differentially private synthetic data, compared to the large-scale benchmark datasets. This provides an opportunity to examine the feasibility of DP-NF in generating useful privacy-preserving synthetic data for small-sized medical datasets with heterogeneous data types and a great degree of realism. We run prediction tasks and variational inference (VI) of a physics-based model on synthetic data generated by DP-NF and compare the utility with that obtained from the original data.

## Preliminaries

NF is defined as a map $F : \mathbb{R}^d \times \mathbf{\Lambda} \to \mathbb{R}^d$ parameterized by $\boldsymbol{\lambda} \in \mathbf{\Lambda}$ transforming realizations from an easy-to-sample base distribution such as $\boldsymbol{z}_0 \sim \mathcal{N}(\mathbf{0}, I_d)$, to realizations from a desired target density. $F$ consists of a composition of $K$ bijections $F_k : \mathbb{R}^d \times \Lambda_k \to \mathbb{R}^d$, each parameterized by $\boldsymbol{\lambda}_k \in \Lambda_k$: $F_{\boldsymbol{\lambda}}(\boldsymbol{z}_0) = F(\boldsymbol{z}_0; \boldsymbol{\lambda}) = [F_K(\,\cdot\,; \boldsymbol{\lambda}_K) \circ F_{K-1}(\,\cdot\,; \boldsymbol{\lambda}_{K-1}) \circ \cdots \circ F_1(\,\cdot\,; \boldsymbol{\lambda}_1)]\,(\boldsymbol{z}_0)$, where $\boldsymbol{z}_k = F_k(\boldsymbol{z}_{k-1}; \boldsymbol{\lambda}_k)$ for $k = 1, \ldots, K$. The distribution of $\boldsymbol{z}_k$, can be obtained by the change of variable technique

$$q_k(\boldsymbol{z}_k) = q_{k-1}(\boldsymbol{z}_{k-1}) \left| \det \frac{\partial F_k}{\partial \boldsymbol{z}_{k-1}} \right|^{-1}. \quad (1)$$

When NF is employed for density estimation given observed data $\boldsymbol{x}$, the following log-likelihood is maximized, assuming $n$ independent observations $\boldsymbol{x}_i$ for $i = 1, \ldots, n$.

$$\ell(\boldsymbol{\lambda}; \boldsymbol{x}) = \log q_K(\boldsymbol{x}) = \textstyle\sum_{i=1}^n \log q_K(\boldsymbol{x}_i)$$
$$= \textstyle\sum_{i=1}^n \log q_0(\boldsymbol{z}_{i,0}) - \sum_{i=1}^n \sum_{k=1}^K \left| \det \frac{\partial F_k}{\partial \boldsymbol{z}_{i,k-1}} \right|, \quad (2)$$

where $\boldsymbol{z}_{i,k-1} = F_k^{-1} \circ \cdots \circ F_{K-1}^{-1} \circ F_K^{-1}(\boldsymbol{x}_i)$. Once the NF parameters are obtained via the maximum likelihood approach, samples from $q_K$ can be generated by first sampling from the base distribution and then transforming the samples through a sequence of bijections.

Among the various NF formulations proposed in the literature, we use masked autoregressive flow (MAF) (Papamakarios, Pavlakou, and Murray 2017). The autoregressive property of MAF is obtained by setting $p(z_i|z_1, \ldots, z_{i-1}) = \phi((z_i - \mu_i)/e^{\alpha_i}$, where $\phi$ is the density function of the standard normal distribution, $\mu_i = f_{\mu_i}(z_1, \ldots, z_{i-1})$, $\alpha_i = f_{\alpha_i}(z_1, \ldots, z_{i-1})$, and $f_{\mu_i}$ and $f_{\alpha_i}$ are masked autoencoder neural networks (MADE) (Germain et al. 2015).

**Definition 1 ($(\epsilon, \delta)$-DP)** *(Dwork et al. 2006b,a) A randomized mechanism $\mathcal{M}$ is $\epsilon$-DP, if $\Pr[\mathcal{M}(D_1) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D_2) \in S] + \delta$ for any $S \in \mathrm{range}(\mathcal{M})$ and neighbouring datasets $D_1, D_2$ that differ by one record, where $\epsilon > 0$ and $\delta \in [0, 1)$.*

DP is a mathematical framework designed to produce robust privacy guarantees. The quantities $\epsilon$ and $\delta$ represent privacy budget or privacy loss parameters. The smaller the $\epsilon$ for a given $\delta$, the higher the privacy protection there is for the individuals in the data.

When DP mechanisms are applied repeatedly to query a dataset, privacy loss will accumulate during the process. It is critical to track the accumulated loss. Abadi et al. (2016) propose a widely popular moment account technique based on Rényi DP (Mironov 2017) to track privacy loss in stochastic gradient descent-based optimization. Dong, Roth, and Su (2022) provide a tighter bound based on Gaussian DP (GDP), which is employed in this work.

**Definition 2 ($\mu$-GDP)** *(Dong, Roth, and Su 2022) Let $T$ be a trade-off function that maps $[0, 1]$ to $[0, 1]$, defined as $T(P_0, P_1)(\alpha) = \inf\{\beta_\phi : \alpha_\phi \leq \alpha\}$ for two distributions $P_1$ from $P_0$ to be distinguished through hypothesis testing $H_0$:$P_0$ vs. $H_1$:$P_1$, where $\phi \in [0, 1]$ is a rejection rule with the type I error $\alpha_\phi = \mathbb{E}_{P_0}[\phi]$ and type II error $\beta_\phi = 1 - \mathbb{E}_{P_1}[\phi]$. A randomized mechanism $\mathcal{M}$ is of $\mu$-GDP if*

$$T(\mathcal{M}(D_1), \mathcal{M}(D_2)) \geq T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1)) \quad (3)$$

*for neighboring datasets $D_1, D_2$ differing by one record.*

The Gaussian mechanism can be applied to achieve $\mu$-GDP guarantees when releasing query results from a dataset (Dong, Roth, and Su 2022). Given a query $f$ to data $D$, a Gaussian mechanism of $\mu$-GDP is $f^*(D) \sim \mathcal{N}(f(D), \Delta_f^2/\mu^2)$, where $\Delta_f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$ for any pair of neighboring datasets $(D_1, D_2)$.

There is a duality between $(\epsilon, \delta)$-DP and $\mu$-GDP (Dong, Roth, and Su 2022), which can be used to translate privacy loss between the two DP frameworks. If a mechanism is of $\mu$-GDP, then it also satisfies $(\epsilon, \delta(\epsilon))$-DP, where $\delta(\epsilon) = \Phi(-\frac{\epsilon}{\mu} + \frac{\mu}{2}) - e^\epsilon \Phi(-\frac{\epsilon}{\mu} - \frac{\mu}{2}))$, where $\Phi$ is the CDF of the standard normal distribution.

## Privacy-preserving Density Estimation and Synthetic Data Generation via DP-NF

In this section, we present differentially private density estimation via NF, which can be used to synthesize privacy-preserving surrogate datasets to release. Users can conduct learning tasks on the synthetic data in the same way as if they had the original data. In what follows, we first present the EHR data that motivates the work, then provide an algorithm to incorporate DP in NF for density estimation, apply the algorithm to generate private EHR data, and examine the utility of the private data in downstream learning tasks including prediction and variational inference (VI).

### Description of the EHR Dataset

The EHR dataset examined in this work was collected in a research project funded by Google through its ATAP initiative, focusing on modeling noninvasive measurements of cardiovascular dynamics. It contains clinical hemodynamic measurements from 82 adult patients suffering from group II pulmonary hypertension (Simonneau et al. 2013) and heart failure with preserved ejection fraction (Harrod et al. 2021). There are 26 attributes measuring heart rate, different types of blood pressure measures, and other cardiovascular-related measures. This dataset focuses on group II pulmonary hypertension, where a reversible increase of pulmonary artery

pressure is caused by an increase in left ventricular filling pressure. An often-used binary classification in practice is based on the diastolic pulmonary artery pressure ($p_{p,d}$) and the systolic artery pressure ($p_{p,s}$), i.e. a patient is hypertensive if the mean pulmonary artery pressure $p_{p,m} = \frac{2}{3}p_{p,d} + \frac{1}{3}p_{p,s} > 20$ mmHg, and not hypertensive otherwise.

## Privacy-preserving Synthetic Data Generation

The dataset, like many EHR datasets, contains missing values. Before synthetic data generation, we first imputed the missing values using the `MICE` package (v3.13.0) in R (Van Buuren and Groothuis-Oudshoorn 2011). Seven attributes have large fractions of missing values (63% to 100%), which turned out to be problematic for the imputation task. For that reason, these 7 attributes were not used in the imputation and any subsequent analysis. We obtained 5 sets of imputed data and then split each of the 5 imputed datasets into a training set and a testing set with 62 and 20 patients, respectively.

We use the general DP-NF procedure in Algorithm 1 for density estimation and generate synthetic data $\boldsymbol{x}^*$ from privacy-preserving density $p^*(\boldsymbol{x})$ in each of the 5 imputed datasets. The algorithm is based on the noisy stochastic gradient descent (SGD) algorithm of $\mu$-GDP in Dong, Roth, and Su (2022) with repeated Gaussian mechanism applications. Denote the number of iterations of an SGD algorithm by $T$, the sub-sampling rate by $r$, and the scale of noise added to the gradient at each iteration by $\sigma$, then the total privacy loss under GDP over $T$ iterations is $\mu = r\sqrt{T(e^{1/\sigma^2} - 1)}$.

The time complexity of Algorithm 1 is $O(Tdn^2r^2) = O(Tnr) + O(Td(nr)^2) + O(Td)$, where the first term is the expected time complexity of the Bernoulli sampling step, the second and third terms are the time complexity of SGD and DP noise sampling from a Gaussian distribution over $T$ interactions, and $d$ is the number of parameters in NF $F_{\boldsymbol{\lambda}}(\cdot)$. Hence Algorithm 1 has the same time complexity as an SGD-based NF algorithm without DP. Though the algorithm is presented using SGD, it can easily be extended to other gradient-based optimization such as RMSprop (Tieleman, Hinton et al. 2012) and Adam (Kingma and Ba 2014), by replacing $\boldsymbol{\lambda}^{*(t+1)} = \boldsymbol{\lambda}^{*(t)} - \eta\, g^{*(t)}$ with their respective parameter update paradigms. Both procedures, with or without DP, have the same time complexity as SGD.

In the application of Algorithm 1 to the EHR data, we used a MAF where the MADE has 15 blocks and a fully connected NN within each block with 1 hidden layer of 200 nodes and the ReLU activation function within and between blocks, resulting in $d = 778,140$ parameters in total. We set $r = 0.5, \eta = 2 \times 10^{-5}, T = 8000, C = 10$; It took about 3.5 hours for the algorithm to generate one synthetic EHR data with $n = 62$ and $p = 19$ attributes. We examined 4 privacy loss settings of $\mu$-GDP at $\mu = (6.10, 3.92, 2.45, 1.49)$, which corresponds to $\sigma = (7.36, 11.44, 18.28, 29.93)$ and $\epsilon = (32, 16, 8, 4)$ if $\delta = 0.01$ in the $(\epsilon, \delta)$-DP setting. We set $r = 0.5$ as larger $r$ would not benefit much from the privacy amplification effect due to subsampling, and smaller $r$ would lead to less stable estimates per iteration. We also attempted different specifications for other hyperparameters, such as $C = 1 \sim 20$ and $\mu \leq 1$. The utility of the synthetic data was either worse or the DP-NF algorithm did not converge due to the large noise injected into the gradients.

Algorithm 1: DP-NF for Density Estimation

**Require:** data set $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, initial parameter values $\boldsymbol{\lambda}^{*(0)}$ of NF $F_{\boldsymbol{\lambda}}(\cdot)$ with base distribution $q_0$, learning rate $\eta$, DP noise scale $\sigma$, subsampling rate $r$, clipping constant on gradient $C$, number of iterations $T$.
**Ensure:** privacy-preserving estimated density $p^*(\boldsymbol{x})$
**for** $t = 0, \cdots, T$ **do**
  sub-sample $\boldsymbol{x}^{(t)}$ from $\boldsymbol{x}$ with rate $r$; let $b^{(t)} = |\boldsymbol{x}^{(t)}|$
  **for** $\boldsymbol{x}_i \in \boldsymbol{x}^{(t)}$ **do**
    $\boldsymbol{z}_i = F_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$
    $l_i = -\log q_0(\boldsymbol{z}_i) - \log|\partial F_{\boldsymbol{\lambda}}/\partial \boldsymbol{z}_i|$
    $\boldsymbol{g}_i^{(t)} = \nabla_{\boldsymbol{\lambda}} l_i$
    $\boldsymbol{g}_i^{(t)} \leftarrow \boldsymbol{g}_i^{(t)} / \max(1, \|\boldsymbol{g}_i^{(t)}\|_2 / C)$
  **end for**
  $\boldsymbol{g}^{*(t)} = \left(\sum_{i=1}^{b^{(t)}} \boldsymbol{g}_i^{(t)} + \mathcal{N}(\boldsymbol{0}, \sigma^2 C^2 \boldsymbol{I})\right) / b^{(t)}$
  $\boldsymbol{\lambda}^{*(t+1)} = \boldsymbol{\lambda}^{*(t)} - \eta\, g^{*(t)}$
**end for**

For each of the 5 imputed training sets, we run DP-NF 10 times to obtain 10 privacy-preserving densities and 10 synthetic datasets were generated from each, leading to a total of 500 synthetic data sets (we generated 500 sets to evaluate the stability of the results in the downstream learning tasks based on one released synthetic dataset. If statistical inferences are of primary interest, one may release multiple synthetic data so as to take into account the uncertainty around sanitization and synthesis, and then combine the inferences from multiple sets using the rule in (Liu 2022)).

We use the generated synthetic data in three subsequent learning tasks. The first two focus on detecting pulmonary hypertension and predicting the mean pulmonary arterial pressure in the testing data while the last conducts VI for the parameters of a physics-based model, presented below.

## Pulmonary Hypertension Prediction Based on Privacy-preserving Synthetic EHR Data

The goal of this learning task is to detect pulmonary hypertension. We train a binary classifier using support vector machines (SVMs) and predict $p_{p,m}$ using random forests (RFs). We adopted the Fowlkes–Mallows (FM) index (Fowlkes and Mallows 1983) to find an optimal cutoff on the predicted probability of hypertension from the SVM rather than using the naïve 0.5 cutoff. Table 1 summarizes the prediction results from the 500 synthetic datasets. As expected, the higher the privacy loss (larger $\mu$), the higher the prediction accuracy and the smaller the prediction MSE.

## VI of Parameters in Physics-based Model Using Privacy-preserving Synthetic EHR Data

The evolution of blood pressure, flow, and volume in the circulatory system of an adult subject is simulated, in this study, using CVSim-6 (Davis 1991; Heldt et al.

| imp. | original | $\mu = \infty$ | $\mu = 6.10$ | $\mu = 2.45$ | $\mu = 1.49$ |
|---|---|---|---|---|---|
| mean classification accuracy rate (SD) via SVM | | | | | |
| 1 | 1.00 | 0.95 (0.07) | 0.91 (0.08) | 0.82 (0.24) | 0.87 (0.09) |
| 2 | 1.00 | 0.88 (0.24) | 0.93 (0.06) | 0.86 (0.10) | 0.83 (0.11) |
| 3 | 1.00 | 0.92 (0.09) | 0.91 (0.06) | 0.90 (0.08) | 0.72 (0.38) |
| 4 | 0.95 | 0.89 (0.10) | 0.88 (0.05) | 0.80 (0.09) | 0.86 (0.12) |
| 5 | 0.95 | 0.85 (0.20) | 0.90 (0.07) | 0.88 (0.17) | 0.80 (0.10) |
| mean | 0.98 | 0.90 (0.16) | 0.91 (0.07) | 0.85 (0.15) | 0.81 (0.20) |
| mean squared error (SD) via RF | | | | | |
| 1 | 20.63 | 40.30 (23.37) | 54.71 (20.20) | 62.45 (20.55) | 70.39 (30.52) |
| 2 | 18.02 | 35.70 (19.87) | 48.08 (18.10) | 60.06 (22.89) | 80.30 (44.03) |
| 3 | 11.50 | 71.76 (62.20) | 36.53 (7.39) | 47.02 (23.01) | 71.36 (30.02) |
| 4 | 18.87 | 32.92 (27.36) | 44.15 (12.19) | 63.09 (19.02) | 60.52 (17.59) |
| 5 | 20.66 | 81.45 (59.57) | 53.98 (17.75) | 59.53 (29.41) | 122.91 (41.70) |
| mean | 17.94 | 52.52 (47.18) | 47.40 (17.17) | 58.41 (23.86) | 80.10 (39.92) |

For the prediction task, we divided the original EHR data into a training dataset of size 64 and a testing dataset of size 20, after imputation. Synthesis data were generated for the training data only, on which SVM and RF were trained.

Table 1: Privacy-preserving prediction (testing data) based on synthetic data generated via DP-NF

2010), a lumped parameter hemodynamic model with 6-compartments available through the PhysioNet repository (Goldberger et al. 2000). The model includes compartments for the left heart, right heart, systemic arteries, systemic veins, pulmonary arteries, and pulmonary veins. Arterial and venous compartments consist of resistance-capacitance (RC) circuits, while heart chambers are simulated using a pressure generator, two unidirectional diodes, and a resistor to account for the pressure loss produced at the valves. All elements are considered linear, disregarding collapsibility in veins associated with negative pressures. Inertial effects are also assumed negligible (the model does not contain inductors) and a two-chamber heart is considered, disregarding any contribution from the atria. In addition, interaction is only considered between adjacent compartments. CVSim-6 consists of a system of six differential equations (one per compartment). The flows between the compartments, under the assumption of nonlinear unidirectional valves (without regurgitation), are expressed as

$$q_{li} = \begin{cases} (P_{pv} - P_l)/R_{li} & \text{if } P_{pv} > P_l \\ 0 & \text{otherwise} \end{cases} \quad q_{lo} = \begin{cases} (P_l - P_a)/R_{lo} & \text{if } P_l > P_a \\ 0 & \text{otherwise} \end{cases}$$

$$q_{ri} = \begin{cases} (P_v - P_r)/R_{ri} & \text{if } P_v > P_r \\ 0 & \text{otherwise} \end{cases} \quad q_{ro} = \begin{cases} (P_r - P_{pa})/R_{ro} & \text{if } P_r > P_{pa} \\ 0 & \text{otherwise} \end{cases}$$

$$q_a = (P_a - P_v)/R_a, \qquad q_{pv} = (P_{pa} - P_{pv})/R_{pv}, \quad (4)$$

leading to the following differential equations for RC elements with fixed and time-varying capacitance

$$\frac{dP_l}{dt} = \frac{q_{li} - q_{lo} - (P_l - P_{th})\, dC_l(t)/dt}{C_l(t)}, \quad \frac{dP_a}{dt} = \frac{q_{lo} - q_a}{C_a},$$
$$\frac{dP_v}{dt} = \frac{q_a - q_{ri}}{C_v}, \quad \frac{dP_r}{dt} = \frac{q_{ri} - q_{ro} - (P_r - P_{th})\, dC_r(t)/dt}{C_r(t)},$$
$$\frac{dP_{pa}}{dt} = \frac{q_{ro} - q_{pv}}{C_{pa}}, \quad \frac{dP_{pv}}{dt} = \frac{q_{pv} - q_{li}}{C_{pv}}. \quad (5)$$

The CVSim-6 model is represented as $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z})$, where $\boldsymbol{z}$ and $\boldsymbol{x}$ represent the model inputs parameters (either unknown or known constants) and outputs (observed), respectively. A subset consisting of 8 outputs $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,8})$ was used in this study, including heart rate, pulmonary vascular resistance, central venous pressure, right ventricular diastolic pressure, right ventricular systolic pressure, right ventricle end-diastolic pressure, average pressure gradient across the aortic valve, and peak pressure gradient across the aortic valve. We regard right ventricular resistance $R_{ro}$ and arterial capacitance $C_a$ as unknown parameters (inputs $\boldsymbol{z}$ to the model) for which statistical inferences are to be obtained given the EHR data, whereas the other parameters are fixed to their default values.

The likelihood functions of $(R_{ro}, C_a)$ given a subset of 43 hypertensive patients in the original and synthetic datasets are assumed to be Gaussian, with independence among the 8 outputs and known marginal variance for each output $l(R_{ro}, C_a; \boldsymbol{x}_i) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{f}(R_{ro}, C_a))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{f}(R_{ro}, C_a))^2\right\}$, where $\boldsymbol{f}(R_{ro}, C_a)$ represents the mean output of the CVSim-6 model and $\boldsymbol{\Sigma}$ is a diagonal matrix with marginal SD (21.1, 12.1, 1.5, 212.2, 21.8, 5.0, 9.2, 3.9) for the 8-dimensional output.

The CVSim-6 model is computationally expensive. To reduce the computational cost when obtaining inference for $R_{ro}$ and $C_a$, we trained offline a fully connected NN surrogate (Wang, Liu, and Schiavazzi 2022). The NN contains 2 hidden layers with 64 and 32 neurons, respectively, and Tanh activations. We generated 100 realizations using Sobol' sampling in the ($R_{ro}$,$C_a$) plane and computed the corresponding outputs from the CVSim-6 model.

Given that $R_{ro}$ and $C_a$ are bounded, we applied the following transformation $R_{ro} = \tanh(3R'/7)(H_R - L_R)/2 + R_0$ and $C_a = \tanh(3C'/7)(H_C - L_C)/2 + C_0$ to obtain unbounded $(R', C')$, where $[L_R, H_R]$, $[L_C, H_C]$ are the respective bounds for $R_{ro}$ and $C_a$ and $R_0, C_0$ are constants and set at 800 and $5 \times 10^{-4}$ in this case. We re-expressed the likelihood function in $(R', C')$ and assumed a uniform prior on $(R', C')$. Given the trained surrogate model, we ran NF to obtain the variational distribution on the transformed $(R', C')$ first, which were then transformed back to obtain the variational posterior distribution of $(R_{ro}, C_a)$.

VI was performed via MAF with 5 alternated MADE and batch normalization layers. Each MADE uses a fully connected NN with 1 hidden layer and 100 neurons. The ReLu activation function was used within and between layers.

The VI results for $R_{ro}$ and $C_a$ based on the privacy-preserving synthetic data are presented in Table 2. The estimates based on privacy-preserving synthetic data are somewhat different from those based on the original data and there does not appear to be an obvious trend over $\mu$ in terms of how the former differs from the latter, except for the pos-

| posterior est | | original | $\mu = 6.10$ | $\mu = 2.45$ | $\mu = 1.49$ |
|---|---|---|---|---|---|
| $R_{ro}$ | mean | 35.7 (1.92) | 43.8 (2.80) | 47.5 (4.40) | 45.6 (3.75) |
| | SD | 1.45 (0.04) | 1.91 (0.03) | 2.19 (0.38) | 2.14 (0.16) |
| $C_a$ | mean | 6.52 (0.13) | 8.57 (1.17) | 7.97 (0.97) | 8.92 (1.02) |
| ($\times 10^{-4}$) | SD | 0.35 (0.01) | 0.61 (0.14) | 0.59 (0.20) | 0.72 (0.19) |
| $R_{ro}, C_a$ | corr. | -0.12 (0.04) | -0.12 (0.19) | -0.21 (0.24) | -0.19 (0.16) |

Table 2: Privacy-preserving VI for CVSim-6 model parameters $R_{ro}$ and $C_a$ given synthetic data generated via DP-NF

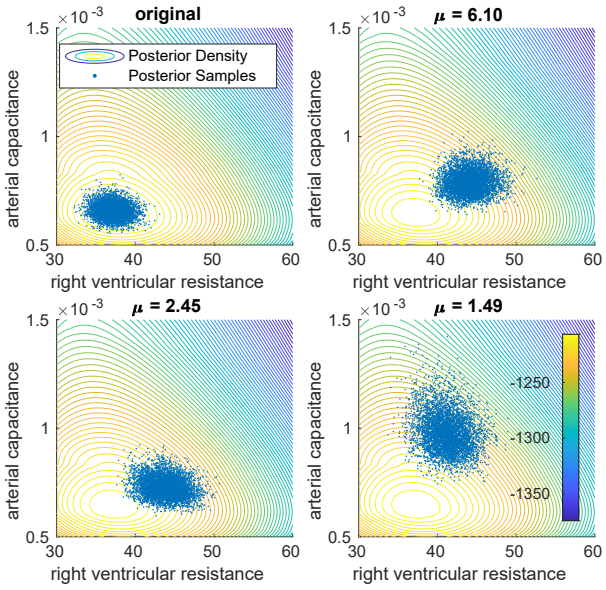


Figure 1: Example of privacy-preserving posterior samples of $(R_{ro}, C_a)$ via NF for VI given one synthetic dataset. The contours represent the densities of the posterior distribution of $(R_{ro}, C_a)$ given the true model and the original data.

terior correlation between $R_{ro}$ and $C_a$, which appears to get more negative as $\mu$, the privacy loss, decreases.

Figure 1 presents the scatter plots of the posterior samples on $R_{ro}$ and $C_a$ from the NF run on one private synthetic dataset. The observations are consistent with those in Table 2. In general, the posterior samples generated from the NF for VI given the privacy-preserving synthetic data are located near the high-density regions of the posterior distribution based on the original data and true model, though not quite around the mode, and the mild negative correlation between the two parameters is also roughly retained for $\mu \geq 2.45$. There is more dispersion among the posterior samples at all $\mu$, especially at $\mu = 1.49$, due to the noise injected to achieve DP guarantees.

# Discussion

## Conclusions

We examined the feasibility and utility of generating synthetic data via privacy-preserving NF for density estimation, with an application to a real EHR dataset. The application focuses on combining inference and computational models to predict the physiology of a given patient at a specific point in time. This is consistent with the choice of a simple hemodynamic model, which offers an accurate characterization of the short-term cardiovascular response.

Overall, the experiments suggest that downstream learning (prediction) and inference (VI) based on privacy-preserving synthetic data yield results of acceptable utility at practically reasonable choices of privacy loss parameters. Despite the pilot nature of the work, we examined a rather difficult case for synthetic data generation with DP via NF, where $p = O(n)$, $n$ is small, and the attributes are of mixed types and have missing values. We conjecture the utility of privacy-preserving synthetic data via the DP-NF procedure would not be worse in datasets with a larger $n$ or $p = o(n)$ compared to the EHR dataset used in this work. Our work contributes to the literature and open-source code on using deep generative models for synthetic medical and healthcare data generation. Furthermore, we examine statistical inference, in particular, VI, in addition to ML tasks based on generated synthetic data, which is as important as, if not more important than, training ML models using synthetic data.

## Limitations and Future Work

This work focuses on NF to generate synthetic data at a single time point with DP guarantees and applies the technique to a tabular EHR dataset with numerical attributes. Built upon the encouraging results from this pilot study, we plan to extend the work in several future directions. Future work will generate more evidence regarding the robustness and generalizability of the DP-NF framework for synthetic EHR data generation and help users to choose which generative models to use when synthesizing EHR data.

First, we plan to compare the computational costs and utility of the synthetic data between the DP-NF procedure with other types of generative models with DP guarantees, such as those listed in the Introduction section and provided by the openDP software (openDP 2003) (MWEM, QUAIL, DP-CTGAN, PATE-CTGAN, and PATE-GAN). A reviewer suggested the procedures in openDP can be slow when dealing with data of large $p$. GANs are known to be difficult to train, especially in small datasets, not to mention with additional DP guarantees. We also had experience running the MWEM procedure (Hardt, Ligett, and McSherry 2012) in other settings (Eugenio and Liu 2021) and found it is not effective at generating continuous data and the quality of synthetic data is sensitive to the number of iterations of the MWEM procedures.

Second, EHR data often contain rich longitudinal information and provide a valuable source for learning temporal trends, predicting future clinical events, and modeling disease progression. It is thus practically important to develop techniques to generate and share synthetic longitudinal and time-series EHR data to make the best of it for these learning tasks. There is already work in the direction either with (Esteban, Hyland, and Rätsch 2017) or without DP guarantees (Yoon, Jarrett, and Van der Schaar 2019; Li et al. 2023) in the GANs framework. For NF, the generation of synthetic time-series EHR data could be achieved based on conditioned normalizing flow (Winkler et al. 2019; Rasul et al. 2020) and Fourier flow (Alaa, Chan, and van der Schaar 2021). As for incorporating DP guarantees, we expect the noisy SGD framework similar to Algorithm 1 still apply, given it is a general framework to achieve DP guarantees through algorithmic perturbation, and is independent of ML model architectures. But the challenges lie in how to best preserve the utility of synthetic data given the temporal correlation and high-dimensionality of the time-series data at an acceptable privacy loss.

Third, we plan to run more empirical studies in more large-scale EHR datasets, in terms of both $n$ and $p$, and

EHR data with more modalities such as text data and imaging data, and other types of unstructured data, in addition to numerical data. Similar to the case of time-series and longitudinal EHR data, developing DP generative models that can synthesize EHR data with multi-modalities is critical for making the best of EHR data for real-world applications. Work exists on this topic, such as Shi et al. (2020); Suzuki and Matsuo (2022) provides a survey on some recent generative models for multi-modal data. Similar to time-series EHR synthetic generation with DP guarantees, the challenges lie in how to best preserve the utility of the multi-modal synthetic data given its complexity and high dimensionality at an acceptable privacy loss.

Lastly, this work is applied in nature. It would be desirable to provide some theoretical guarantees to the utility of the noisy SGD algorithm with DP guarantees. This is an active research topic in general, especially for non-convex optimization.

## Code

A custom Python/Cython implementation of the physics-basaed CVSim model can be found at *https://github.com/desResLab/supplMatHarrod20*. The code for density estimation through DP-NF is available at https://github.com/cedricwangyu/DPNF. Users may use the code to reproduce the results in the current study or apply the code to other datasets where the goal is either DP-NF for density estimation or synthetic data generation.

## Acknowledgments

This work is supported by NSF grant #1918692 (PI DES), NSF CAREER grant #1942662 (PI DES), NSF grant #2104831 (Notre Dame PI DES), and used computational resources provided through the Center for Research Computing at the University of Notre Dame, USA.

The EHR dataset utilized in this work contains data from external studies that involved human participants and were conducted in compliance with ethical standards. This study is classified as research not involving human subjects and was approved on June 13th, 2019, by the Office of Research Compliance and Institutional Review Board at the University of Notre Dame under IRB#19-05-5371.

## References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Alaa, A.; Chan, A. J.; and van der Schaar, M. 2021. Generative time-series modeling with Fourier flows. In *International Conference on Learning Representations*.

Beaulieu-Jones, B. K.; Wu, Z. S.; Williams, C.; Lee, R.; Bhavnani, S. P.; Byrd, J. B.; and Greene, C. S. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7): e005122.

Benaim, A. R.; Almog, R.; Gorelik, Y.; Hochberg, I.; Nassar, L.; Mashiach, T.; Khamaisi, M.; Lurie, Y.; Azzam, Z. S.; Khoury, J.; et al. 2020. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR medical informatics*, 8(2): e16492.

Bowen, C. M.; and Liu, F. 2020. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2): 280–307.

Bu, Z.; Dong, J.; Long, Q.; and Su, W. J. 2020. Deep learning with Gaussian differential privacy. *Harvard data science review*, 2020(23).

Chen, Q.; Xiang, C.; Xue, M.; Li, B.; Borisov, N.; Kaarfar, D.; and Zhu, H. 2018. Differentially private data generative models. *arXiv preprint arXiv:1812.02274*.

Chen, R. J.; Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; and Mahmood, F. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6): 493–497.

Dash, S.; Dutta, R.; Guyon, I.; Pavao, A.; Bennett, K. P.; et al. 2019. Privacy preserving synthetic health data. In *2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Davis, T. L. 1991. *Teaching physiology through interactive simulation of hemodynamics*. Ph.D. thesis, Massachusetts Institute of Technology.

Dong, J.; Roth, A.; and Su, W. J. 2022. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1): 3–37.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, 486–503. Springer.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Esteban, C.; Hyland, S. L.; and Rätsch, G. 2017. Real-valued (medical) time series generation with recurrent conditionalGANs. *arXiv preprint arXiv:1706.02633*.

Eugenio, E. C.; and Liu, F. 2021. Construction of Differentially Private Empirical Distributions from a Low-Order Marginals Set Through Solving Linear Equations with l 2 Regularization. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 3*, 949–966. Springer.

Fowlkes, E. B.; and Mallows, C. L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383): 553–569.

Germain, M.; Gregor, K.; Murray, I.; and Larochelle, H. 2015. MADE: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, 881–889. PMLR.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hardt, M.; Ligett, K.; and McSherry, F. 2012. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25.

Harrod, K. K.; Rogers, J. L.; Feinstein, J. A.; Marsden, A. L.; and Schiavazzi, D. E. 2021. Predictive Modeling of Secondary Pulmonary Hypertension in Left Ventricular Diastolic Dysfunction. *Frontiers in Physiology*, 12.

Heldt, T.; Mukkamala, R.; Moody, G. B.; and Mark, R. G. 2010. CVSim: an open-source cardiovascular simulator for teaching and research. *The open pacing, electrophysiology & therapy journal*, 3: 45.

Jordon, J.; Yoon, J.; and Van Der Schaar, M. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Lee, J.; Kim, M.; Jeong, Y.; and Ro, Y. 2022. Differentially Private Normalizing Flows for Synthetic Tabular Data Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36 (7), 7345–7353.

Li, J.; Cairns, B. J.; Li, J.; and Zhu, T. 2023. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digital Medicine*, 6(1): 98.

Liu, F. 2022. Model-based differentially private data synthesis and Statistical Inference in Multiply Synthetic Differentially Private Data. *Transactions on Data Privacy*, 15(3): 141–175.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275. IEEE.

openDP. 2003. smartnoise-synth. https://opendp.org/contribute. [Online; accessed 17-June-2023].

Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.

Pfitzner, B.; and Arnrich, B. 2022. DPD-fVAE: Synthetic Data Generation Using Federated Variational Autoencoders With Differentially-Private Decoder. *arXiv preprint arXiv:2211.11591*.

Rankin, D.; Black, M.; Bond, R.; Wallace, J.; Mulvenna, M.; Epelde, G.; et al. 2020. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics*, 8(7): e18910.

Rasul, K.; Sheikh, A.-S.; Schuster, I.; Bergmann, U.; and Vollgraf, R. 2020. Multivariate probabilistic time series forecasting via conditioned normalizing flows. *arXiv preprint arXiv:2002.06103*.

Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International conference on machine learning*, 1530–1538. PMLR.

Rubin, D. B. 1993. Statistical disclosure limitation. *Journal of official Statistics*, 9(2): 461–468.

Shi, Y.; Paige, B.; Torr, P. H.; and Siddharth, N. 2020. Relating by contrasting: A data-efficient framework for multimodal generative models. *arXiv preprint arXiv:2007.01179*.

Simonneau, G.; Gatzoulis, M.; Adatia, I.; Celermajer, C., D.and Denton; Ghofrani, A.; Sanchez Gomez, M.; Kumar, R.; Landzberg, M.; Machado, R.; Olschewski, H.; Robbins, I.; and R., S. 2013. Updated clinical classification of pulmonary hypertension. *Journal of the American College of Cardiology*, 62(25 Supplement): D34–D41.

Suzuki, M.; and Matsuo, Y. 2022. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6): 261–278.

the SHARED team. 2023. The Synthetic Health And Research Data (SHARED) Project. Accessed on Jan 8, 2023.

Tieleman, T.; Hinton, G.; et al. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26–31.

Van Buuren, S.; and Groothuis-Oudshoorn, K. 2011. MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45: 1–67.

Waites, C.; and Cummings, R. 2021. Differentially private normalizing flows for privacy-preserving density estimation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 1000–1009.

Wang, Y.; Liu, F.; and Schiavazzi, D. E. 2022. Variational inference with NoFAS: Normalizing flow with adaptive surrogate for computationally expensive models. *Journal of Computational Physics*, 467: 111454.

Winkler, C.; Worrall, D.; Hoogeboom, E.; and Welling, M. 2019. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Xie, L.; Lin, K.; Wang, S.; Wang, F.; and Zhou, J. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*.

Yale, A.; Dash, S.; Dutta, R.; Guyon, I.; Pavao, A.; and Bennett, K. 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416: 244–255.

Yoon, J.; Jarrett, D.; and Van der Schaar, M. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32.