Privacy-preserving generative deep neural networks support clinical data sharing

Authors: Brett K. Beaulieu-Jones[1], Zhiwei Steven Wu[2], Chris Williams[3], James Brian Byrd[4], Casey S. Greene[3]*

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[2]Computer and Information Science, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[3]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[4]Division of Cardiovascular Medicine, Department of Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA.

*To whom correspondence should be addressed: csgreene@upenn.edu

One Sentence Summary: Deep neural networks can generate shareable biomedical data to allow reanalysis while preserving the privacy of study participants.

Abstract: Though it is widely recognized that data sharing enables faster scientific progress, the sensible need to protect participant privacy hampers this practice in medicine. We train deep neural networks that generate synthetic subjects closely resembling study participants. Using the SPRINT trial as an example, we show that machine-learning models built from simulated participants generalize to the original dataset. We incorporate differential privacy, which offers strong guarantees on the likelihood that a subject could be identified as a member of the trial. Investigators who have compiled a dataset can use our method to provide a freely accessible public version that enables other scientists to perform discovery-oriented analyses. Generated data can be released alongside analytical code to enable fully reproducible workflows, even when privacy is a concern. By addressing data sharing challenges, deep neural networks can facilitate the rigorous and reproducible investigation of clinical datasets.

Introduction

Sharing individual-level data from clinical studies remains challenging. The status quo often requires scientists to establish a formal collaboration and execute extensive data usage agreements before sharing data. These requirements slow or even prevent data sharing between researchers in all but the closest collaborations.

Recent initiatives have begun to address cultural challenges around data sharing. The New England Journal of Medicine recently held the Systolic Blood Pressure Trial (SPRINT) Data Analysis Challenge to examine possible benefits of clinical trial data sharing *(1, 2)*. The SPRINT clinical trial examined the efficacy of intensive lowering of systolic blood pressure (<120 mmHg) compared with treatment to a standard systolic blood pressure goal (<140 mmHg). Intensive blood pressure lowering resulted in fewer cardiovascular events and the trial was stopped early. Reanalysis of the challenge data led to the development of personalized treatment scores *(3)* and decision support systems *(4)*, in addition to a more specific analysis of blood pressure management in participants with chronic kidney disease *(5)*. Initiatives such as the SPRINT Data Analysis Challenge have begun to address cultural norms. Even for this effort which focused on data sharing, investigators were required to execute data use agreements that included clauses to maintain security and prohibit re-identification or sharing.

We sought to alleviate privacy barriers that hamper data sharing. Park and Ghosh developed an initial approach to managing privacy threats using a perturbed Gibbs sampler to generate synthetic data with a quantifiable privacy risk *(6)*. Goodfellow et al. *(7)* developed a method entitled Generative Adversarial Networks (GANs) using neural networks to generate realistic samples from complex distributions. GANs have become a class of widely used machine learning methods and have recently been used in biology and medicine *(8)*. In this work, we trained two deep neural networks against each other to generate realistic simulated participant blood pressure trajectories and medication adjustments from the SPRINT trial dataset. One neural network, called the generator, is trained to generate a participant from a set of random numbers. The other neural network, called a discriminator, is trained to classify data as real or generated. As the networks are trained, the generator learns to build samples that fool the discriminator. Networks trained in this way are called GANs and can also be used for labeled samples *(9)*. A pair of recent preprints have reported participant generation via neural networks *(10, 11)*. Of particular note, Esteban et al., evaluated their synthetic samples in a transfer learning task, training a machine learning algorithm on one dataset and applying it to another. Esteban et al. trained classifiers on synthetic data and then tested them on real data to evaluate whether the synthetic samples are realistic enough. However, it is not enough to simply build synthetic participants. Numerous linkage and membership inference attacks have demonstrated the ability to re-identify participants or reveal participation in a study on both biomedical datasets *(12–19)* and from machine learning models *(20–22)*.

To provide a formal privacy guarantee, we build GANs under the constraint of differential privacy (16). Differential privacy protects against common privacy attacks including membership inference, homogeneity and background knowledge attacks *(23)*. Informally, differential privacy requires that no subject in the study has a significant influence on the information released by the algorithm (see Materials and Methods for a formal definition). Despite being a stringent notion, differential privacy allows us to generate new plausible individuals while revealing almost nothing about any single study participant. Within the biomedical domain, Simmons and Berger developed a method using differential privacy to enable privacy preserving genome-wide association studies *(24)*. Recently, methods have also been developed to train deep neural networks under differential privacy with formal assurances about privacy risks *(25, 26)*. In the context of a GAN, the discriminator is the only component that accesses the real, private, data. By training the discriminator under differential privacy, we can produce a differentially private GAN framework.

We evaluated whether or not this approach could generate biomedical data that could be shared for reanalysis while reducing participant privacy risks. We evaluated usefulness by: (1) comparing variable distributions between the real and simulated data, (2) comparing the correlation structure between variables in the real and simulated data, (3) comparing machine learning predictors constructed on real vs. simulated data. We find that the model learns realistic distributions and that models constructed from the simulated data successfully classify participants in a held-out portion of the underlying real dataset.

Results

We used an Auxiliary Classifier Generative Adversarial Network (AC-GAN) *(9)* to simulate participants based on the population of the SPRINT clinical trial. We included all participants with measurements for the first twelve time periods (n=6,502), dividing them into a training set (n=6,000) and a test set (n=502). We trained two AC-GANs using the training set: a traditional, standard, AC-GAN (results labeled non-private throughout the remainder of this manuscript) and an AC-GAN trained under differentially privacy (results labeled private). We used both GANs to simulate data that we compared to the real data. We visualized participant blood pressure trajectories, analyzed variable correlation structure and evaluated transfer learning performance for a machine learning classification task. In addition, we had a cardiologist attempt to classify participants as real or synthetic and whether they were in the standard or intensive treatment group.

Auxiliary Classifier GAN for SPRINT Clinical Trial Data.

An AC-GAN (Fig. 1A) is made up of two neural networks competing with each other. We found convolutional layers effectively modeled the sequential measurements and used deep convolutional neural networks for both the generator and discriminator (Fig. 1B, 1C). We trained the Generator (G) to take in a specified treatment arm (standard/intensive) and random noise and generate new participants that can fool the Discriminator (D). The generator simulated a systolic blood pressure, diastolic blood pressure and a number of medications for each synthetic patient. We trained the discriminator to differentiate real and simulated data from a dataset containing both groups. We repeated this process until the generator created synthetic participants that were difficult to discriminate from real ones.

We trained under differential privacy by limiting the effect any single subject has on the training process and by adding random noise based on the maximum effect of a single subject. From the technical perspective, we limited the effect of participants by clipping the norm of the discriminator's training gradient and added proportionate Gaussian noise. This combination ensures that training cannot be tied to an individual and that it could have been guided by a different subject within or outside the real training data. The maximum effect of an outlier is limited and bounded. Comparing the neural network loss functions of the private and non-private training process demonstrates the effects of these constraints. Under normal training the losses of the generator and discriminator converged to an equilibrium before eventually increasing steadily (Fig. 1D). Under differentially private training the losses converged to and remained in a noisy equilibrium (Fig. 1F). At the beginning of training the neural networks changed rapidly. As training continued and the model achieved a better fit these steps, the gradient, decreased. When the gradient became very small, the noise outweighed the signal and limited further training.

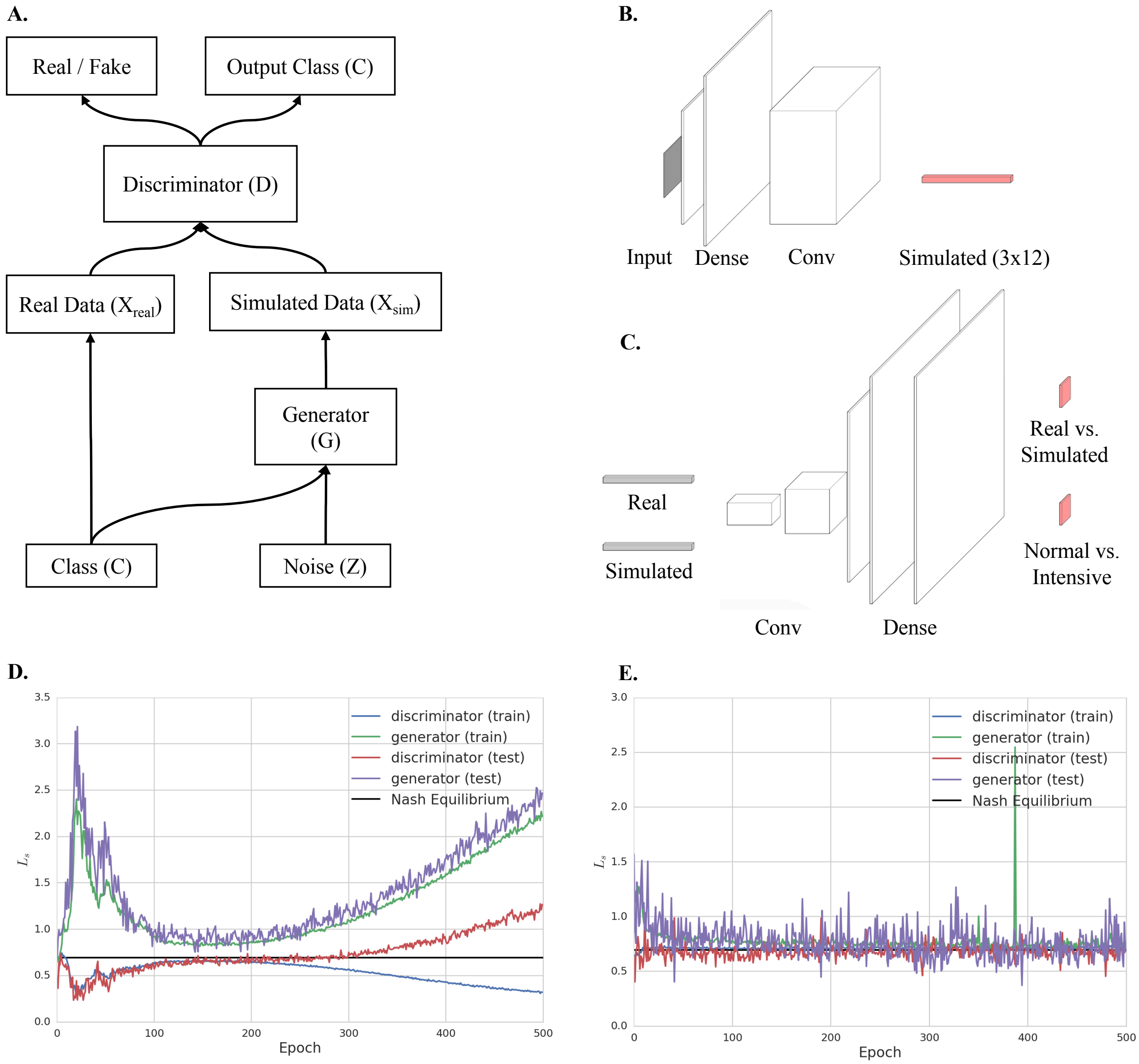


Fig. 1. AC-GAN architecture and training. **A.)** Structure of an AC-GAN. **B.)** The generator model takes a class label and random noise as input and outputs a 3x12 vector for each participant (SBP, DBP and medication counts at each time point). **C.)** The discriminator model takes both real and simulated samples as input and learns to predict the source and a class label (i.e. normal or intensive treatment group). **D.)** Training loss for a non-private AC-GAN. **E.)** Training loss for a private AC-GAN.

Evaluation of Simulated Participants

After training the AC-GAN we compared the simulated synthetic participants to the real participants (Figure 2). Figure 2 shows the median systolic blood pressures for: (1) real participants, (2) simulated participants via a non-private AC-GAN and (3) simulated participants via the differentially private AC-GAN. The non-private participants generated at the end of training appear similar to the real participants. The private participants have wider variability because of the noise added during training (Fig. 2A). As the models achieve better fit, the gradient shrinks, causing the gradient to noise ratio to decrease. This can occasionally lead to the private generator and discriminator falling out of sync (Supp. Fig. 1) or more commonly the private model generating less realistic samples due to noise. To best select epochs where synthetic samples closely real samples, we tested each epoch's data by training an additional classifier that must distinguish whether a generated participant was a part of the normal or intensive treatment groups. We applied two common machine learning classification algorithms and selected the top epochs in a differentially private manner (Fig. 2B and 2C).

However, selecting only a single epoch does not account for the AC-GAN training process. Because the discriminator and generator compete from epoch to epoch, their results can cycle around the underlying distribution. The non-private models consistently improved throughout training (Supp. Fig. 2A, Supp. Fig. 3A), but this could be due to the generator eventually learning characteristics specific to individual participants. We observed that epoch selection based on the training data was important for the generation of realistic populations from models that incorporated differential privacy (Supp. Fig. 2B, Supp. Fig. 3B). To address this, we simulated 1,000 participants from each of the top five epochs by both the logistic regression and random forest evaluation on the training data and combined them to form a multi-epoch training set. This process maintained differential privacy and resulted in a generated population that, throughout the trial, was consistent with the real population (Fig. 2D). The epoch selection process was independent of the holdout testing data.

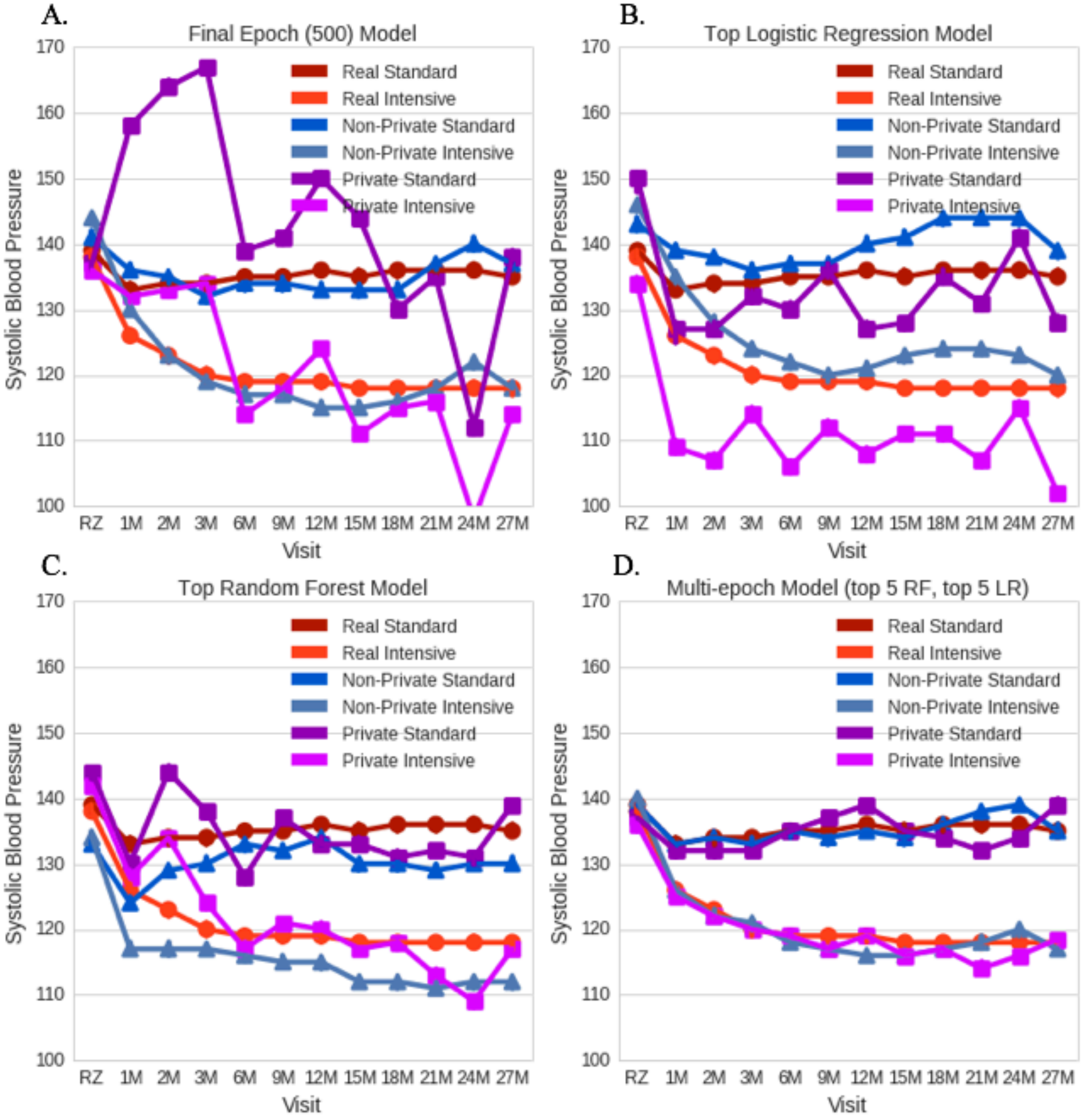


Fig. 2. Median Systolic Blood Pressure Trajectories from initial visit to 27 months. **A.)** Simulated samples (private and non-private) generated from the final (500th) epoch of training. **B.)** Simulated samples generated from the epoch with the best performing logistic regression classifier. **C.)** Simulated

samples from the epoch with the best performing random forest classifier. **D.)** Simulated samples from the top five random forest classifier epochs and top five logistic regression classifier epochs.

The Pearson correlation structure of the real data (Fig. 3A) was closely reflected by the correlation structure of the non-private generated data (Fig. 3B). Of note was initial positive correlation between the number of medications a participant was taking and the early systolic blood pressures, but this correlation decreased as time goes on. The correlation matrices between the real training data and the non-private data were highly correlated (Spearman correlation = 0.9645, p-value < 10^{-325}). The private generated data generally reflects these trends, but has an increased level of noise (Fig. 3C). The correlation matrices between the real training data and the private generated data were only slightly less correlated (Spearman correlation = 0.8787, p-value = 7.692^{-204}. The noisy training process of the private discriminator places an upper bound on its ability to fit the distribution of data. Increased sample sizes would help to clarify this distribution and because larger sample sizes cause less privacy loss, less noise would need to be added to achieve an acceptable privacy budget.

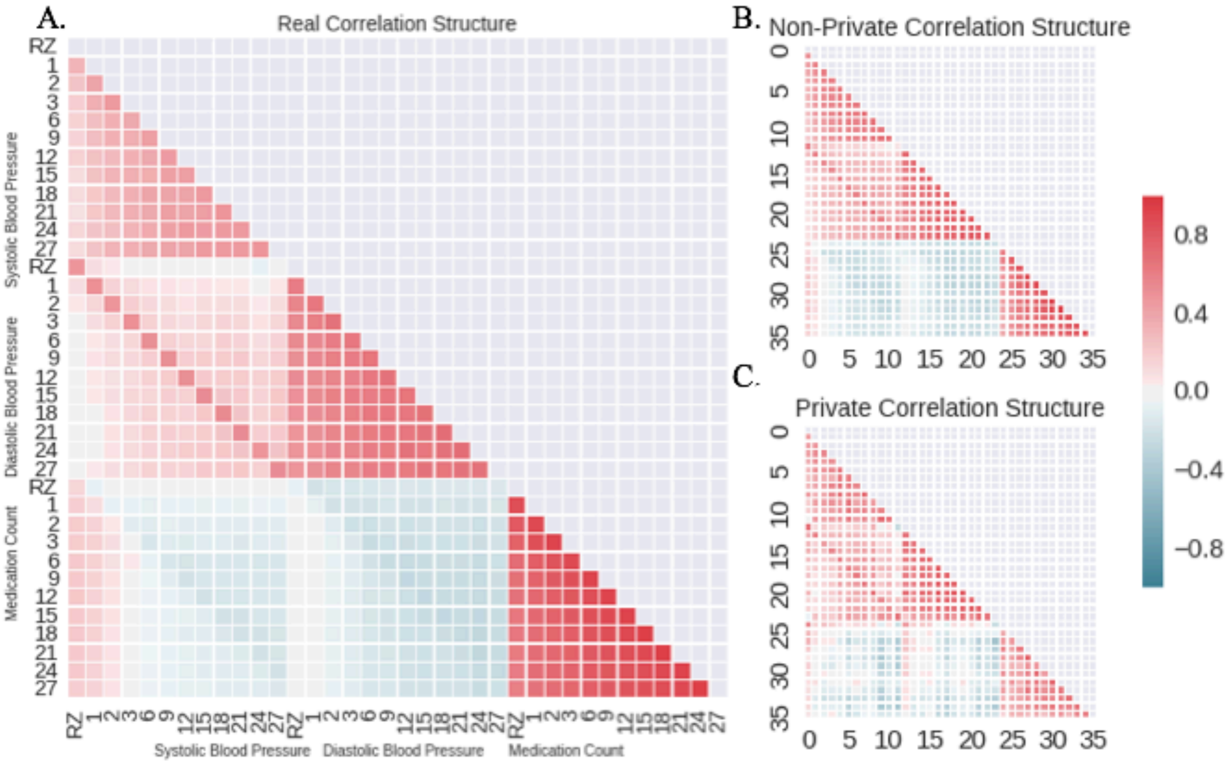


Fig. 3. Pairwise Pearson correlation between columns for the **A.)** Original, real data, **B.)** Non-private, AC-GAN simulated data **C.)** Differentially private, AC-GAN simulated data.

To determine whether the synthetic patients looked realistic to an experienced clinician, we presented 100 participants (50 real, 50 synthetic) to a cardiologist who is certified by the American Society of Hypertension as a hypertension specialist, and who had carefully reviewed the SPRINT trial inclusion criteria and study protocol. The cardiologist looked for data inconsistent with the SPRINT protocol or that otherwise appeared anomalous. For example, the clinician was alert for instances in which the systolic blood pressure was less than 100 mm Hg, but the participant was prescribed an additional medication. The cardiologist classified each record on a zero to ten realism scale (10 was the most realistic) , as well as whether the data correspond to standard or intensive treatment (Fig. 4A-D). The mean realism score for synthetic patients was 4.52 and the mean score for the real patients 5.08 (Figure 4E). We performed a Mann-Whitney U test to evaluate whether the scores were drawn from significantly different distributions and found a low but not significant p-value (0.0825). The cardiologist correctly classified 43 (86%) of both the real and synthetic patients as the standard or intensive group.

Feasibility of Simulated Participants for Transfer Learning Task

Clinician review, visualizations of patient distributions and variable correlations showed that synthetic participants appeared similar to real participants. We sought to determine whether or not synthetic participants could be used for subsequent data mining. We trained machine learning classifiers using four methods (logistic regression, random forests, support vector machines, and nearest neighbors) to distinguish treatment groups on three different sources of data: real participants, synthetic participants generated by the non-private model, and synthetic participants generated by the private model. We compared performance of these classifiers on a holdout test set of 502 real participants (Fig. 5 A-D). This analysis revealed two main trends: classifiers trained on the set constructed from combined top epochs exhibited more stable performance on the test data in line with observations from the population distributions, and classifiers trained on data from the non-private model slightly outperformed those trained on data from the private model. A drop in performance was expected because adding noise to maintain privacy reduces signal. If desired, training a non-private model could provide an upper bound for expected performance.

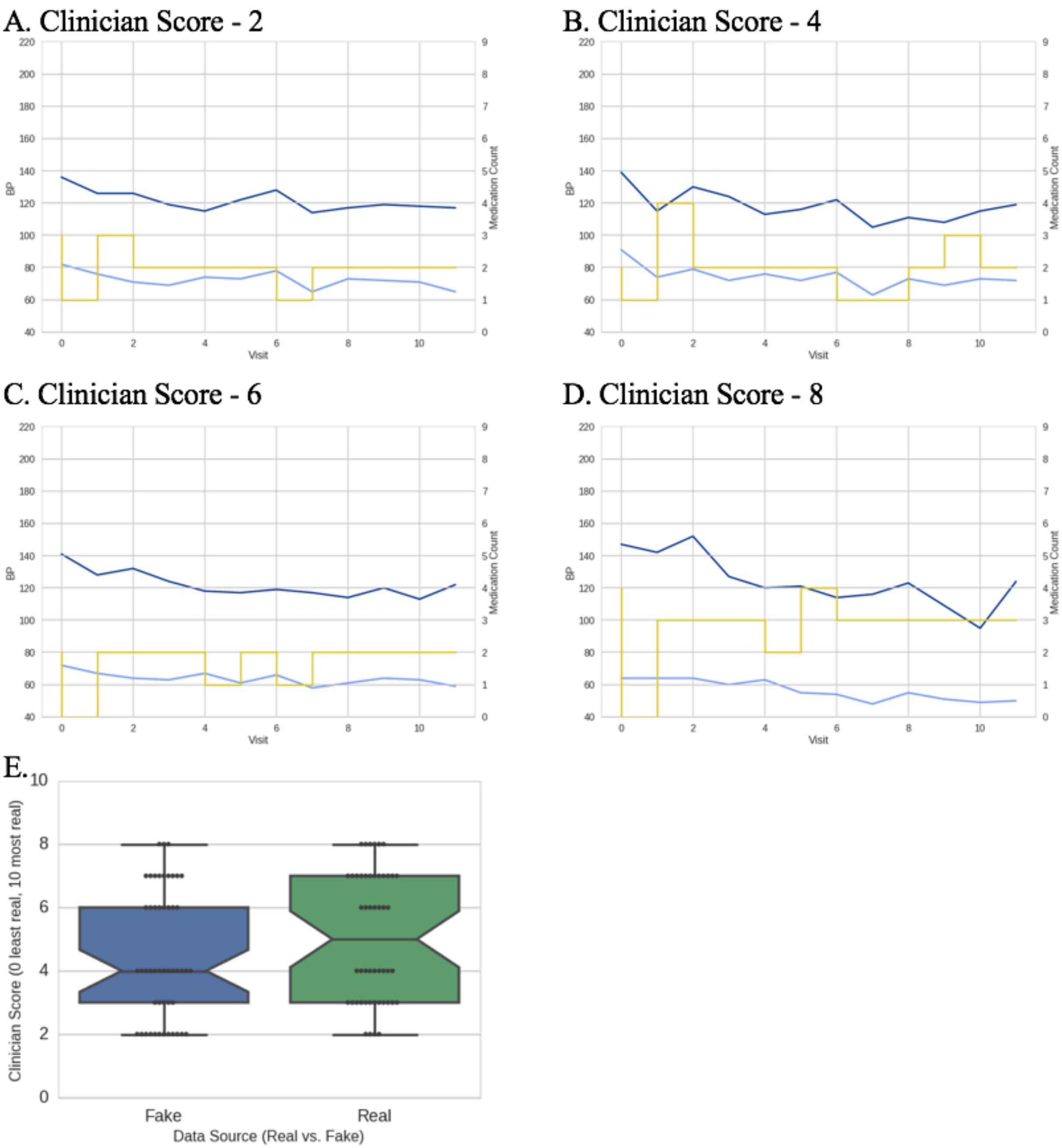


Fig 4. A.) Synthetic patient scored a 2 by clinician expert. **B.)** Synthetic patient scored a 4 by clinician expert. **C.)** Synthetic patient scored a 6 by clinician expert. **D.)** Synthetic patient scored a 8 by clinician expert. **E.)** Comparison of the distribution of scores between real and fake patients.

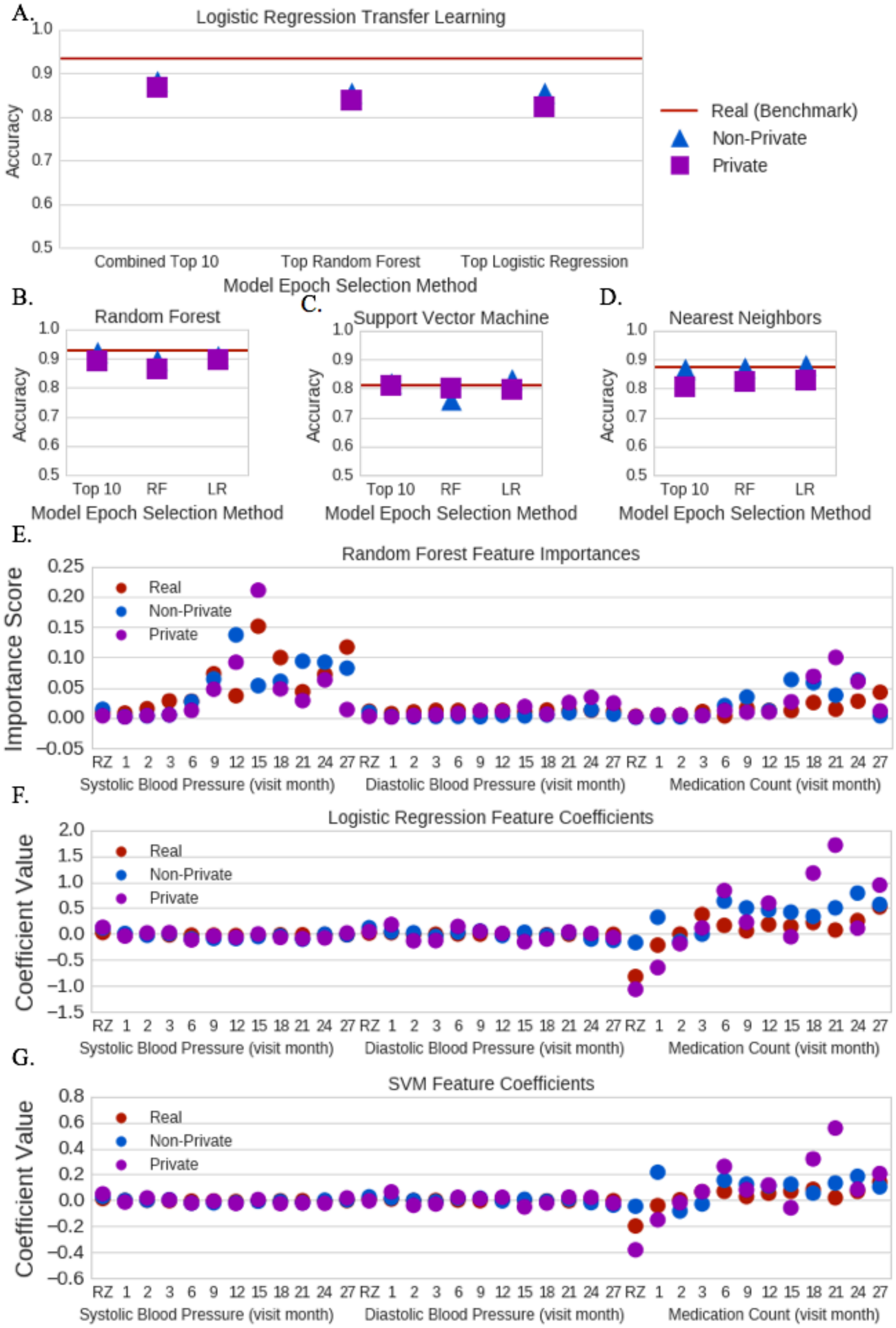

A.
Logistic Regression Transfer Learning
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Combined Top 10
Top Random Forest
Top Logistic Regression
Model Epoch Selection Method
Real (Benchmark)
Non-Private
Private
B.
Random Forest
C.
Support Vector Machine
D.
Nearest Neighbors
Accuracy
Top 10
RF
LR
Model Epoch Selection Method
E.
Random Forest Feature Importances
Importance Score
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
Real
Non-Private
Private
RZ 1 2 3 6 9 12 15 18 21 24 27
Systolic Blood Pressure (visit month)
Diastolic Blood Pressure (visit month)
Medication Count (visit month)
F.
Logistic Regression Feature Coefficients
Coefficient Value
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
G.
SVM Feature Coefficients
Coefficient Value
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Fig 5. A-D.) Performance on transfer learning task by source of training data for each machine learning method. **E.)** Random forest variable importance scores by training data. **F.)** Logistic Regression variable coefficients by training data. **G.)** Support Vector Machine variable coefficients by training data.

We also sought to determine the extent to which the classifiers were using similar predictive features. We evaluated the random forest feature importance scores (Fig. 5E) as well as the logistic regression and support vector machine feature coefficients (Fig. 5G, 5F). All showed similar trends of useful features between real and generated data, and a Spearman correlation test was performed between the importance scores (random forest) and coefficients (SVM and logistic regression) of the models trained on real data and each synthetic set revealed significant associations in all cases (Table 1). Though all three classification methods achieved similar accuracy, the random forest classifier found the medication features to be important while these features had near zero coefficients in the SVM and logistic regression classifiers.

Table 1. Spearman Correlation between variable importance scores (Random Forests) and model coefficients (Support Vector Machine and Logistic Regression).

	Random Forest		**Support Vector Machine**		**Logistic Regression**	
	Correlation	P-Value	Correlation	P-Value	Correlation	P-Value
Real - Non-Private	0.7207	7.1518e-07	0.5279	9.35794e-04	0.6973	2.2950e-06
Real - Private	0.6769	5.7988e-06	0.6895	3.2918e-06	0.6692	8.0932e-06

Privacy Analysis

The formal definition of differential privacy has two parameters. The key parameter ε measures the "privacy loss" incurred by the computation. The second parameter δ bounds the probability that the privacy loss exceeds ε. The values of (ε, δ) accumulate as the algorithm repeatedly accesses the private data. In our experiment, our private AC-GAN algorithm is able to generate useful synthetic data with ε = 2 and $\delta < 10^{-5}$ (Fig. 6). The upper bound of the epoch selection task, (see Materials Methods) used (0.05, 0) per each model included for a total of (0.5, 0) differential privacy. This established a modest, single digit epsilon privacy budget of (2.5, 10^{-5}).

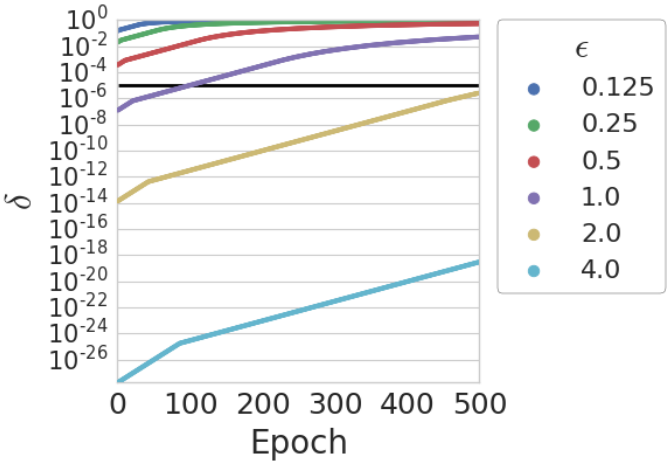


Fig 6. The value of delta as a function of epoch for different epsilon values. An ε value of 2 allows for 500 epochs of training and $\delta < 10^{-5}$.

Predicting Heart Failure in the MIMIC Critical Care Database

We tested whether our approach could be applied in a second dataset by predicting heart failure from the first five measurements for nine vital sign measurements in 7,222 patients from the MIMIC Critical care database. Performance on privately generated synthetic patients was on par with performance models trained on real patients (Fig. 7A-D). As in the SPRINT data, the coefficients for logistic regression and the support vector machine as well as the feature importances were significantly correlated between real and synthetic data (Table 2). Interestingly, the synthetic data showed generally higher inter-feature correlation but these correlations were also highly correlated with the real data (Spearman correlation – 0.5598, p-value – 2.995^{-53}).

Table 2. Spearman Correlation between variable importance scores (Random Forests) and model coefficients (Support Vector Machine and Logistic Regression).

	Random Forest		**Support Vector Machine**		**Logistic Regression**	
	Correlation	P-Value	Correlation	P-Value	Correlation	P-Value
Real - Non-Private	0.4059	0.00566	0.2952	0.04894	0.4268	0.00345

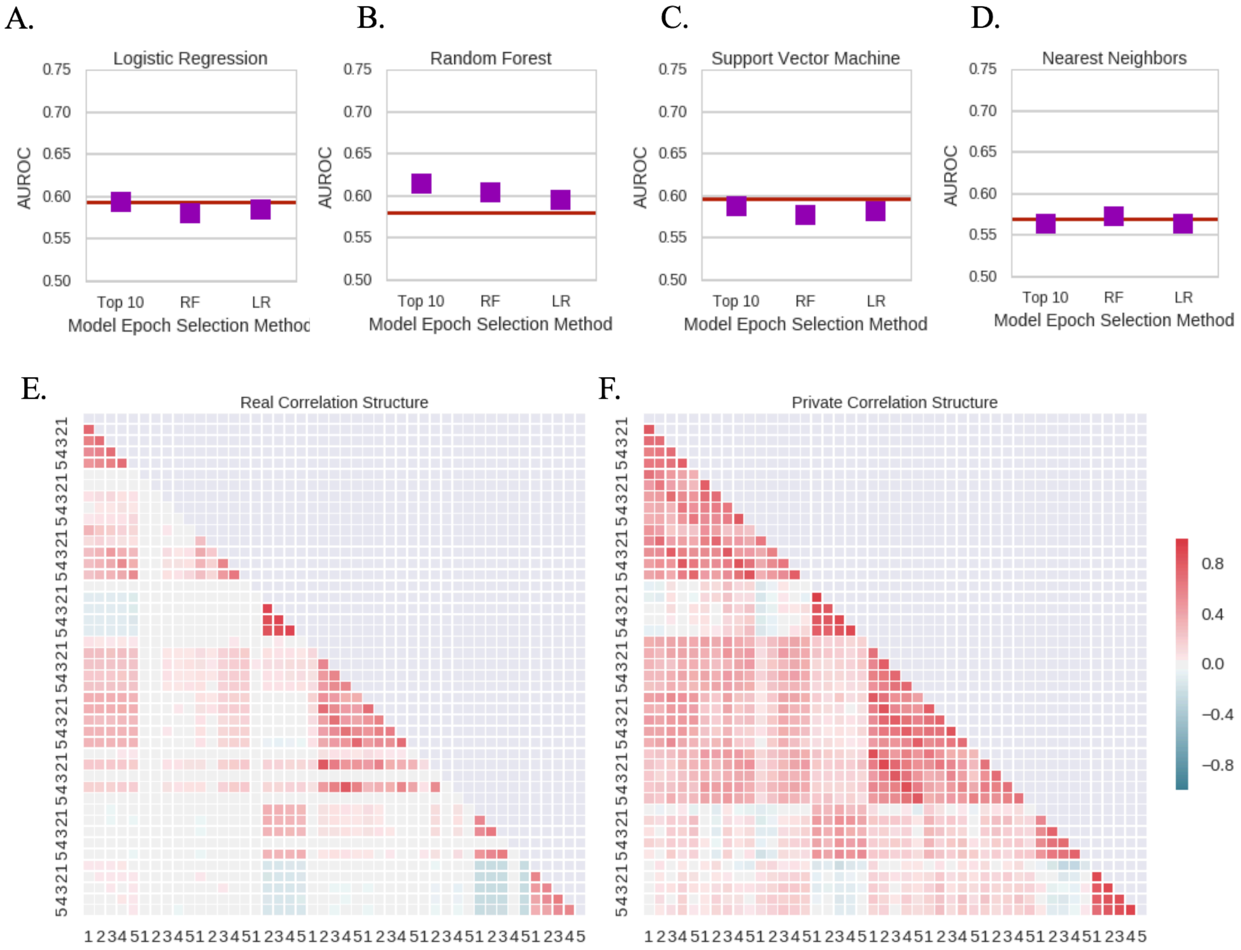


Fig 7. A-D.) Performance on transfer learning task by source of training data for each machine learning method. **E.)** Pairwise Pearson correlation between columns for the Original, real data **F.)** Pairwise Pearson correlation between columns for the Private synthetic data.

Discussion

Deep generative adversarial networks and differential privacy offer a technical solution to the challenge of sharing biomedical data to facilitate exploratory analyses. Our approach, which uses deep neural networks for data simulation, can generate synthetic data to be distributed and used for secondary analysis. We perform training with a differential privacy framework that limits the study subjects' privacy risk. We apply this approach to data from the SPRINT clinical trial due to its recent use for a data reanalysis challenge

We introduce an approach that samples from multiple epochs to improve performance while maintaining privacy. However, several challenges remain. Deep learning models have many training parameters and require substantial sample sizes, which can hamper this method's use for small clinical trials or targeted studies. Another fruitful area of use may be large electronic health records systems, where the ability to share synthetic data may aid methods development and the initial discovery of predictive models. Similarly, financial institutions or other organizations that use outside contractors or consultants to develop risk models might choose to share generated data instead of actual client data. In very large datasets, there is evidence that differential privacy may even prevent overfitting to reduce the error of subsequent predictions *(27)*.

Though our approach provides a general framing, the precise neural network architecture may need to be tuned for specific use cases. Data with multiple types presents a challenge. EHRs contain binary, categorical, ordinal and continuous data. Neural networks require these types to be encoded and normalized, a process that can reduce signal and increase the dimensionality of data. New neural networks have been designed to deal more effectively with discrete data *(28, 29)*. Researchers will need to incorporate these techniques and develop new methods for mixed types if their use case requires it. We expect this approach to be most well suited to sharing specific variables from clinical trials to enable wide sharing of data with similar properties to the actual data. We do not intend the method to be applied to generate high dimensional genetic data from whole genome sequences or other such features. Application to that problem would require the selection of a subset of variants of interest or substantial additional methodological work.

Due to the fluid nature of security and best practices, it is important to choose a method which is mathematically provable and ensures that any outputs are robust to post-processing. Differential privacy satisfies both needs and is thus being relied upon in the upcoming 2020 United States Census *(30)*. It is imperative to remember that to receive the guarantees of differential privacy a proper implementation is required. We believe testing frameworks to ensure accurate implementations are a promising direction for future work, particularly in domains with highly sensitive data like healthcare.

The practice of generating data under differential privacy with deep neural networks offers a technical solution for those who wish to share data to the challenge of patient privacy. This technical work complements ongoing efforts to change the data sharing culture of clinical research.

Materials and Methods

We developed an approach to train auxiliary classifier generative adversarial networks (AC-GANs) in a differentially private manner to enable privacy preserving data sharing. Generative adversarial networks offer the ability to simulate realistic-looking data that closely matches the distribution of the source data. AC-GANs add the ability to generate labeled samples. By training AC-GANs under the differential privacy framework we generated realistic samples that can be used for initial analysis while guaranteeing a specified level of participant privacy.

The source code for all analyses is available under a permissive open source license in our repository (https://github.com/greenelab/SPRINT_gan). In addition, continuous analysis *(31)* was used to re-run all analyses, to generate docker images matching the environment of the original analysis, and to track intermediate results and logs. These artifacts are freely available (https://hub.docker.com/r/brettbj/sprint-gan/ and archival version: https://doi.org/10.6084/m9.figshare.5165731.v1).

SPRINT Clinical Trial Data

The SPRINT was a randomized, single blind treatment trial where participants were randomized into two groups, an intensive treatment group with a systolic blood-pressure target of less than 120 mmHg and a standard treatment group with a systolic blood-pressure target of less than 140 mm Hg. The trial included a total of 9,361 participants. We included 6,502 participants from the trial by filtering for all participants that had blood pressure measurements for each of the first 12 measurements (RZ, 1M, 2M, 3M, 6M, 9M, 12M, 15M, 18M, 21M, 24M, 27M). We included measurements for systolic blood pressure, diastolic blood pressure and the count of medications prescribed to each participant. This provided an input vector of shape (3, 12).

Auxiliary Classifier Generative Adversarial Network

We implemented the AC-GAN as described in Odena et al. *(9)* using Keras *(32)* to simulate systolic and diastolic blood pressures as well as the number of hypertension medications prescribed. Results shown use a latent vector of dimension 100, a learning rate of 0.0002, and a batch size of 1 trained for 500 epochs. To conform with the privacy claims laid out in Abadi et al. *(25)*, gradients must be clipped per example. In our implementation this requires the batch size to be 1. To handle edge cases and mimic the sensitivity of the real data measurements, we take the floor of zero or the simulated value and convert all values to integers. Full implementation details can be seen in the GitHub repository (https://github.com/greenelab/SPRINT_gan/blob/master/ac_gan.py).

Clinician Evaluation

We constructed figures showing blood pressure variables and medication count for each time point. An American Society of Hypertension-certified hypertension specialist made a “real or synthetic” determination for each participant showing systolic blood pressure, diastolic blood pressure, and number of medications at each of 12 visits. The cardiologist classified how realistic the patients looked (from 1-10 where 10 is most realistic) and whether the patients were a part of the standard or intensive treatment plan. Prior to reviewing the figures and regularly during the review of figures, the clinician reviewed the published SPRINT protocol to help contextualize the data. We performed a power analysis using the samplesize R package for ordinal Mann-Whitney U tests. We performed a Mann-Whitney U test to evaluate whether the real or synthetic samples received significantly different scores and compared the accuracy of the treatment plan classifications. With a hypothetical distribution of scores defining the effect size of the synthetic set (0=0.1, 1=0.15, 2=0.2, 3=0.3, 4=0.3, 5=0.2, 6=0.15, 7=0.15, 8=0.1, 9=0.05, 10=0.05) and the mirrored distribution used for the real set, an alpha of 0.05, and half of the samples being from the synthetic set we found that 80 samples would provide 80% power. We provided the physician with 100 samples for evaluation.

Transfer Learning Task

Each of the 6,502 participants in our analytical dataset is labeled by treatment group. We evaluate machine learning methods (logistic regression, support vector machines, and random forests from the scikit-learn *(33)* package) by their ability to predict which group a participant belongs to. This was done by splitting the 6,502 participants into a training set of 6,000 participants (labeled real) and a test set of 502 participants. A vanilla AC-GAN was trained using the 6,000-participant training set providing a simulated training set (labeled non-private). A differentially private AC-GAN was trained using the

6,000-participant training set providing a differentially private training simulated training set (labeled private). Each classifier was then trained on the real, non-private and private training sets and evaluated on the same, real test set of participants. This allows for a comparison of classification performance between models trained on the real data, synthetic data and private synthetic data. We evaluated both accuracy as well as the correlation between important features (random forest) and model coefficients (logistic regression and support vector machine).

Differential Privacy

Differential privacy is a stability property for algorithms, specifically for randomized algorithms *(34)*. Informally, it requires that the change of any single data point in the data set has little influence on the output distribution by the algorithm. To formally define differential privacy, let us consider X as the set of all possible data records in our domain. A dataset is a collection of n data records from X. A pair of datasets D and D' are neighboring if they differ by at most one data record. In the following, we will write R to denote the output range of the algorithm, which in our case correspond to the set of generative models.

Definition 1 [Differential Privacy *(35)*]: Let $\varepsilon, \delta > 0$. An algorithm $A: X^n \rightarrow R$ satisfies (ε, δ)-differential privacy if for any pair of neighboring datasets D, D', and any event $S \subseteq R$, the following holds

$$\Pr[A(D) \in S] \leq \Pr[A(D') \in S] \exp(\varepsilon) + \delta,$$

where the probability is taken over the randomness of the algorithm.

A crucial property of differential privacy is its resilience to post-processing --- any data independent post-processing procedure on the output by a private algorithm remains private. More formally:

Lemma [Resilience to Post-Processing]: Let algorithm $A: X^n \rightarrow R$ be an (ε, δ)-differentially private algorithm. Let $A' : R \rightarrow R'$ be a "post-processing" procedure. Then their composition of running A over the dataset D, and then running A' over the output A(D) also satisfies (ε, δ)-differential privacy.

Training AC-GANs in a Differentially Private Manner

During the training of AC-GAN, the only part that requires direct access to the private (real) data is the training of the discriminator. To achieve differential privacy, we only need to "privatize" the training of the discriminators. The differential privacy guarantee of the entire AC-GAN directly follows because the output generative models are simply post-processing from the discriminator.

To train the discriminator under differential privacy we add noise to the stochastic gradient descent process as outlined in Abadi et al. *(25)*. First, we provide an upper bound onto the norm of the gradient at any individual step. This is done by clipping the ℓ^2-norm of the gradient. Next, we perturb each coordinate of the gradient by adding noise drawn from a Gaussian distribution with a variance proportional to the gradient clipping. The more noise we added (relative to the clipped norm of the gradient) the better privacy guarantee. To achieve a modest privacy budget, we found we could clip the ℓ^2-norm of the gradient at 0.0001 and add noise from a normal distribution with a σ^2 of 1 ($N(\mu, 1 * (0.0001^2))$). This is substantially higher than previously shown, likely due to either the dynamic nature of GAN training where the target is inexact and changes over time or averaging over many mini-batches. We used the moments accountant described in Abadi et al. *(25)* to compute the privacy parameters (ε, δ). These parameters were determined after running a grid search for noise (0.25, 0.5, 1, 1.5, 2, 3, 4, 8) and gradient clipping (0.1, 0.01, 0.001, 0.0001, 0.00001) to determine how long models could be trained under (ε, δ) of $(2.5, 10^{-5})$.

Differentially Private Model Selection

We found that sampling from multiple different epochs throughout training provided a more diverse training set. This provided summary statistics closer to the real data and higher accuracy in the transfer learning task. During the GAN training, we saved all the generative models across all epochs. We then generated a batch of synthetic data from each generative model, and used a machine learning algorithm (logistic regression or random forest) to train a prediction model based on each synthetic batch of data. We then tested each prediction model on the training set from the real dataset and calculate the resulting accuracy. To select epochs that generate training data for the most accurate models under differential privacy, we used the standard "Report Noisy Min" subroutine: first add independent Laplace noise to the accuracy of each model (drawn from Lab(1/(n*ε)) to achieve (ε, 0) differential privacy where n is the size of the private dataset we perform the prediction on and output the model with the best noisy accuracy.

In practice, we choose the top five models that performed best on the transfer learning task for the training data using both logistic regression classification and random forest classification (for a total of 10 models). We performed this task under (0.5, 0)-differential privacy. In each of the ten rounds of selection epsilon was set to 0.05. This achieves a good balance of accuracy while maintaining a reasonable privacy budget.

Predicting Heart Failure in the MIMIC Critical Care Database

We applied the method to the MIMIC Critical Care Database *(36)* to demonstrate its generality. We generated synthetic patients for the purpose of predicting Heart Failure. MIMIC is a database of 46,297 de-identified electronic health records for critical care patients at Beth Israel. We defined patients who suffered from Heart Failure as any patient in MIMIC diagnosed with an ICD-9 code included in the Veterans Affair's Chronic Heart Failure Quality Enhancement Research Initiative's guidelines (402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428, 281.1, 428.20, 428.21, 428.22, 428.23, 428.30, 428.31, 428.32, 428.33, 428.40, 428.41, 428.42, 428.43, and 428.9). We performed complete case analysis for patients with at least five measurements for mean arterial blood pressure, arterial systolic and diastolic blood pressures, beats per minute, respiration rate, peripheral capillary oxygen saturation (SpO2), mean non-invasive blood pressure and mean systolic and diastolic blood pressures. For patients with more than five measurements for these values, the first five were used. This yielded 8,260 total patients and 2,110 cases of heart failure. We included the first 7,500 patients in the training set and the remaining 760 in a hold-out test set. The training and transfer learning procedures matched the SPRINT protocol. Because the classes were unbalanced, we used f1 score to evaluate the results from the transfer learning exercise.

Acknowledgments: We thank Jason H. Moore (University of Pennsylvania), Aaron Roth (University of Pennsylvania), Gregory Way (University of Pennsylvania), Yoseph Barash (University of Pennsylvania), Anupama Jha (University of Pennsylvania) and Blanca Himes (University of Pennsylvania) for their helpful discussions. This Manuscript was prepared using SPRINT_POP Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the SPRINT_POP or the NHLBI. We thank the participants of the SPRINT trial and the entire SPRINT Research Group. **Funding:** This work was supported by the Gordon and Betty Moore Foundation under a Data Driven Discovery Investigator Award to C.S.G. (GBMF 4552). B.K.B.-J. Was supported by a Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health and by US National Institutes of Health grants AI116794 and LM010098. Z.S.W is funded in part by a subcontract on the DARPA Brandeis project and a grant from the Sloan Foundation. J.B.B. is funded by US National Institutes of Health grant K23-HL128909. **Author Contributions:** B.K.B.-J. and C.S.G. conceived the study. B.K.B.-J. And C.W. performed initial analyses. B.K.B.-J. and Z.S.W. designed and validated the privacy approach. J.B.B

performed a blinded review of records. B.K.B.-J., C.S.G. and Z.S.W. wrote the manuscript and all authors revised and approved the final manuscript. **Competing interests:** The authors have no competing interests to disclose. **Data and materials availability:** All data used in this manuscript are available via the NHLBI (https://biolincc.nhlbi.nih.gov/studies/sprint_pop/), the source code is available via GitHub (https://github.com/greenelab/SPRINT_gan) and an archived version is available via Figshare (DOI: 10.6084/m9.figshare.5165737).

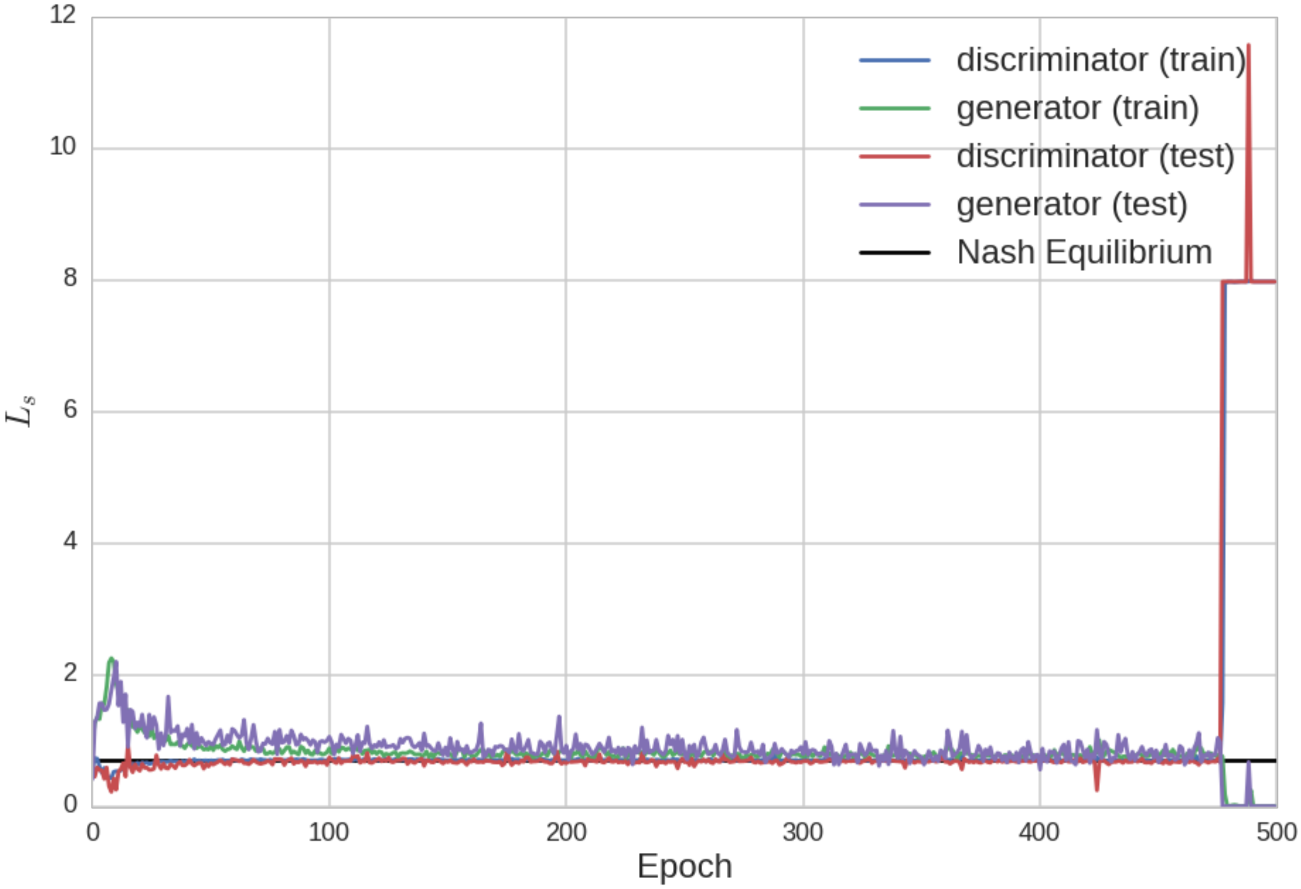
References:

1. The SPRINT Data Analysis Challenge 2017 (available at https://challenge.nejm.org/pages/home).
2. S. R. Group, A randomized trial of intensive versus standard blood-pressure control, *N Engl J Med* (2015) (available at https://www.nejm.org/doi/full/10.1056/NEJMoa1511939).
3. S. Basu, J. B. Sussman, J. Rigdon, L. Steimle, B. Denton, R. Hayward, Development and Validation of a Clinical Decision Score to Maximize Benefit and Minimize Harm from Intensive Blood Pressure Treatment (2017) (available at https://challenge.nejm.org/posts/5815).
4. N. Dagan, M. A. Tsadok, M. Hoshen, A. Arkiv, T. Karpati, I. Gofer, M. Leibowitz, H. Gilutz, E. Podjarny, E. Bachmat, R. Balicer, To Treat Intensively or Not – Individualized Decision Making Support Tool (2017) (available at https://challenge.nejm.org/posts/5826).
5. R. Aggarwal, J. Steinkamp, N. Chiu, M. H. Sang, J. Park, H. Mirzan, B. Petrie, Assessing the Impact of Intensive Blood Pressure Management in Chronic Kidney Disease Patients (2017) (available at https://challenge.nejm.org/posts/5837).
6. Y. Park, J. Ghosh, M. Shankar, in *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, (2013), pp. 493–498.
7. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, 2672–2680 (2014).
8. T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, A. Gitter, C. S. Greene, Opportunities And Obstacles For Deep Learning In Biology And Medicine, *bioRxiv* (2017) (available at https://doi.org/10.1101/142760).
9. A. Odena, C. Olah, J. Shlens, Conditional Image Synthesis With Auxiliary Classifier GANs, (2016) (available at http://arxiv.org/abs/1610.09585).
10. E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, J. Sun, Generating Multi-label Discrete Electronic Health Records using Generative Adversarial Networks, (2017) (available at http://arxiv.org/abs/1703.06490).
11. C. Esteban, S. L. Hyland, G. Rätsch, Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs, (2017) (available at http://arxiv.org/abs/1706.02633).
12. S. L. Garfinkel, De-Identification of Personal Information, , doi:10.6028/NIST.IR.8053.
13. K. El Emam, E. Jonker, L. Arbuckle, B. Malin, J. Riedl, R. W. Scherer, Ed. A Systematic Review of Re-Identification Attacks on Health Data, *PLoS One* **6**, e28071 (2011).
14. B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems, *J. Biomed. Inform.* **37**, 179–192 (2004).
15. I. S. Kohane, R. B. Altman, Health-Information Altruists ? A Potentially Critical Resource, *N. Engl. J. Med.* **353**, 2074–2077 (2005).
16. L. Sweeney, k-anonymity: A model for protecting privacy, *Int. J. Uncertainty, Fuzziness* (2002) (available at http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648).
17. L. Sweeney, A. Abu, J. Winn, Identifying participants in the personal genome project by name, (2013) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2257732).
18. A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, *Privacy, 2008. SP 2008. IEEE* ... (2008) (available at http://ieeexplore.ieee.org/abstract/document/4531148/).
19. N. Homer, S. Szelinger, M. Redman, D. Duggan, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, *PLoS* (2008) (available at http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167).
20. M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, *Proc. 22nd ACM* (2015) (available at http://dl.acm.org/citation.cfm?id=2813677).

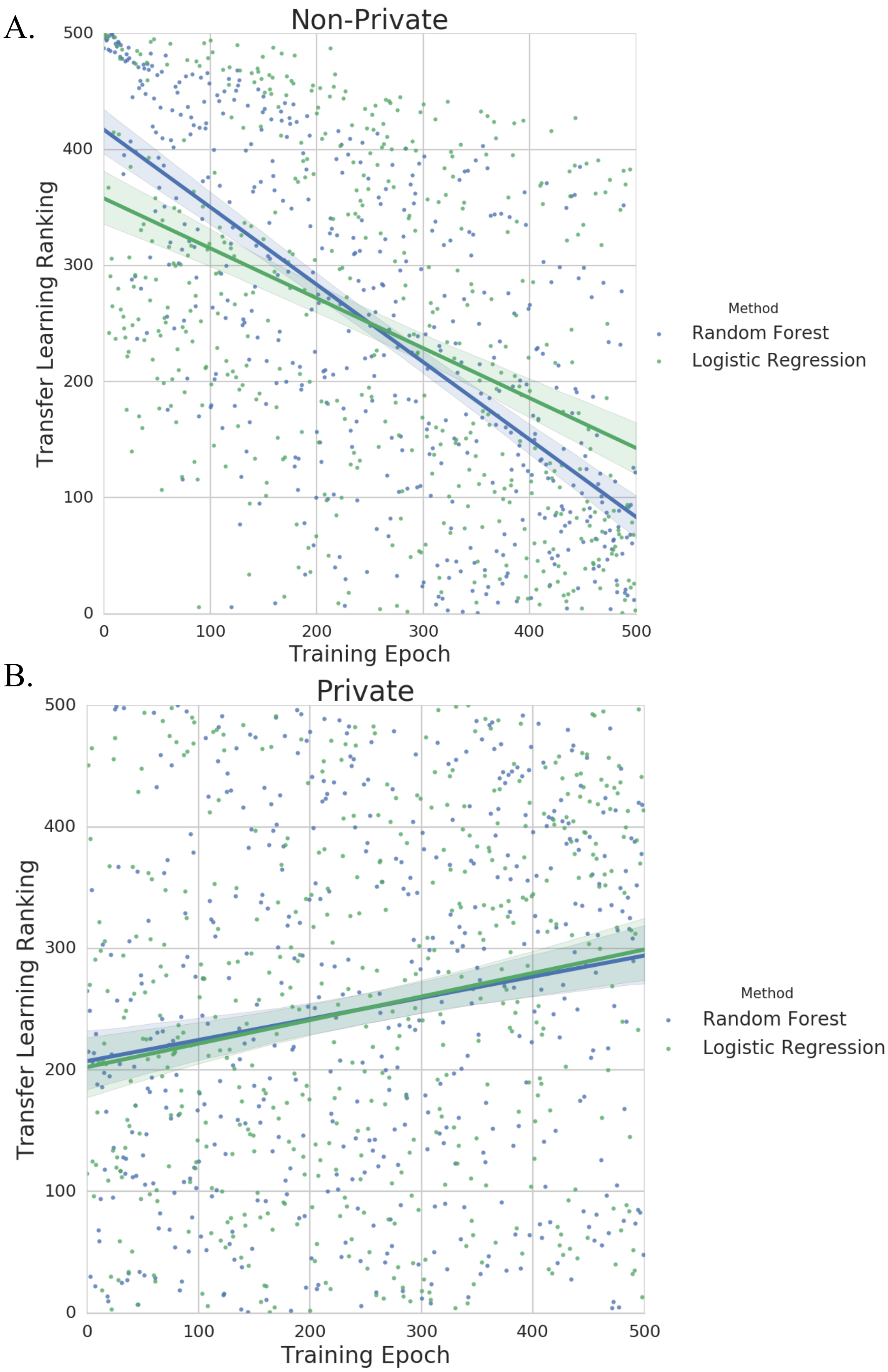
21. R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership Inference Attacks against Machine Learning Models, (2016) (available at http://arxiv.org/abs/1610.05820).
22. M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing., *USENIX* (2014) (available at https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf).
23. C. Dwork, A. Roth, The Algorithmic Foundations of Differential Privacy, *Found. Trends® Theor. Comput. Sci.* **9**, 211–407 (2013).
24. S. Simmons, B. Berger, Realizing privacy preserving genome-wide association studies, *Bioinformatics* (2016) (available at http://bioinformatics.oxfordjournals.org/content/32/9/1293.short).
25. M. Abadi, A. Chu, I. Goodfellow, H. McMahan, Deep learning with differential privacy, *Proc.* (2016) (available at http://dl.acm.org/citation.cfm?id=2978318).
26. R. Shokri, V. Shmatikov, Privacy-preserving deep learning, *Proc. 22nd ACM SIGSAC* (2015) (available at http://dl.acm.org/citation.cfm?id=2813687).
27. C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, A. Roth, The reusable holdout: Preserving validity in adaptive data analysis, *Science (80-.).* **349** (2015) (available at http://science.sciencemag.org/content/349/6248/636).
28. E. Jang, S. Gu, B. Poole, Categorical Reparameterization with Gumbel-Softmax, (2016) (available at http://arxiv.org/abs/1611.01144).
29. M. J. Kusner, J. M. Hernández-Lobato, GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution, (2016) (available at http://arxiv.org/abs/1611.04051).
30. T. Adams, R. Ashmead, A. Dajani, J. Devine, M. Hay, C. Hollingsworth, M. Ibrahimi, M. Ikeda, P. Leclerc, A. Machanavajjhala, C. Martindale, G. Miklau, B. Moran, N. Porter, A. Ross, W. Sexton, Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test, (2017).
31. B. K. B. Beaulieu-Jones, C. C. S. Greene, Reproducibility of computational workflows is automated using continuous analysis, *Nat Biotech* **35**, 342–346 (2017).
32. F. Chollet, *Keras* (GitHub, 2015; http://203.195.193.174/nat123CacheFolder/646F63732E626470742E6E6574/35c3a8a4cb1d4160bfd7b6e93d74ab67CD30CE37D036D032DF31CE3ACC30C533C9_e22880a46b0f1f3f3eb1e14dd5452984/media/pdf/kerascn/latest/kerascn.pdf#page=59).
33. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, *… Mach. Learn. …* **12**, 2825–2830 (2012).
34. D. Cynthia, Differential privacy, *Autom. Lang. Program.* (2006).
35. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, *Theory Cryptogr.* (2006) (available at http://link.springer.com/chapter/10.1007/11681878_14).
36. A. Johnson, T. Pollard, L. Shen, L. Lehman, MIMIC-III, a freely accessible critical care database, *Scientific* (2016) (available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878278/).

Supplemental Materials

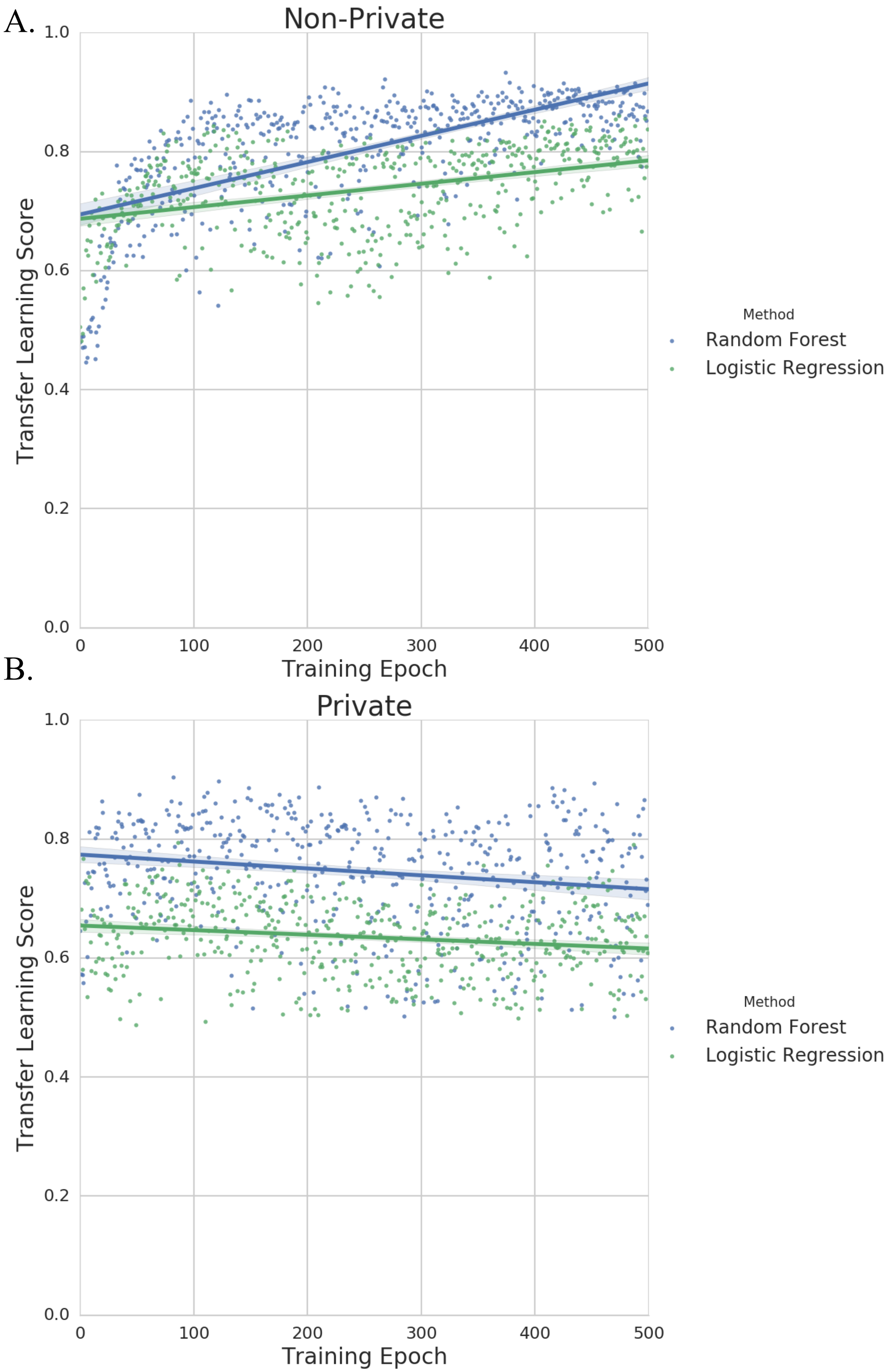
Supplemental Figure 1. Random noise breaks equilibrium.



Supplemental Figure 2. Top Ranking Epochs for Transfer Learning Exercise



Supplemental Figure 3. Scores vs. Epoch for Transfer Learning Task.


A.
Non-Private
Transfer Learning Score
Training Epoch
Method
Random Forest
Logistic Regression
B.
Private
Transfer Learning Score
Training Epoch
Method
Random Forest
Logistic Regression